Study of DNA Physics by a High-Throughput Genome Mapping Technique

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Hui-Min Chuang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisor: **Kevin D. Dorfman**

September, 2019

© Hui-Min Chuang 2019
ALL RIGHTS RESERVED

Acknowledgements

I begin by expressing my utmost gratitude to my advisor, Prof. Kevin Dorfman, for giving me such a great opportunity to join his group and focus on the projects that motivated me to study multiple fields of science to explore my interest. I appreciate the opportunity he provided for me to work collaboratively with Bionano Genomics. Thanks Kevin for his guidance not only for my research but also for my life, and his warm encouragement whenever I felt frustrated with my work. Kevin is the best mentor and advisor I have ever had. Thank you.

I would like to thank my other committee members, Prof. Samira Azarin, Prof. Benjamin Hackel and Prof. Joachim Mueller for reading and commenting on my work. I would also like to thank Aditya Bikram Bhandari, Guo Kang Cheong, Paridhi Agrawal and Zixue Ma for proofreading parts of the thesis.

Thanks to Dr. Jeff Reifenberger for our collaboration and his effort for teaching me experimental skills the alignment of microscope.

I appreciate the greatest support from the people in Dorfman group, Michael McGovern, Vaidyanathan Sethuraman, Akash Arora, Scott White, Pranav Agrawal, Xiaolan Li, Seunghwan Shin, Aditya Bikram Bhandari, Guo Kang Cheong, Mathew Thomas, Paridhi Agrawal and Zixue Ma. I enjoyed every group lunch and happy hour in these years with all laugh and complain together. Your company made every day of my Ph.D. life. Special thanks to Julian Sheats and Damini Gupta who mentored me and taught me valuable research skills when I joined the group in the beginning.

I am really grateful to my CEMS and out-of-department friends, Pin-Kuang Lai, Wenjia Zhang, Ting-Pi Sun, Szu-Ming Yang, Jason Chen, Yao-Chang Yang, and Chih-Jui Lin. Thank you for hanging out with me and all of the support over my most difficult moments during the years.

Finally, I would like to acknowledge financial support from the Government Scholarship for Study Abroad from the Ministry of Education of Taiwan, University of Minnesota's Doctoral Dissertation Fellowship, the National Institutes of Health (R01-HG006851) and the Department of Chemical Engineering & Materials Science. I also learned some fabrication techniques in the Nano Fabrication Center at the University of Minnesota, which receives partial support from the NIH.

Dedication

To my parents, Yu-Kuei Chuang and Hsiao-Chen Hsia, and my sister, Wei-Ting Chuang, who support me over the years.

Abstract

Next generation sequencing (NGS) is a powerful tool to sequence DNA with a low error rate at an affordable cost. While NGS is successful with these features, its short read length leads to difficulty in detecting structural variation of chromosome at a large scale. The technique of labeling barcode can help resolve these drawbacks by generating scaffolds for the detection of large-scale structural variation and *de novo* whole genome mapping in combination with NGS. Though the labeling barcode technique is well-developed experimentally, how correctly interpreting the data remains a challenge. In fact, the behavior of DNA in confinement is not thoroughly understood. There is still disagreement between experiments and theory that cannot be fully explained. The sources for the discrepancy may be an oversimplified model or assumption about the properties of DNA used to interpret the results. The aim of this dissertation research is to correct the assumption, review an existing model, and propose an approach to modify the model.

We first corrected an oversimplified assumption about one of DNA physical properties, the persistence length, in the model we usually used for long DNA. Work on confined DNA usually assumes there is no dependence of the persistence length on DNA sequence when DNA is long enough to average over the intrinsic curvature. We correct this rough assumption using a novel approach to data collection and interpretation, and development of a model to rationalize the experimental results. Using a high-throughput genome mapping technique, we obtained circa 50 million measurements of the extension of internal human DNA segments in a 41 nm $\times$ 41 nm nanochannel. The underlying DNA sequences, obtained by mapping to the reference human genome, reveal that the DNA persistence length increases by almost 20% as the percent of guanine-cytosine (G-C) base pairs increases. The increased persistence length is rationalized by a statistical terpolymer model. This developed model, which contains a sequence-dependent intrinsic persistence length and a sequence-independent electrostatic persistence length, can help predict the persistence length of any long DNA sequence for the analysis of DNA-based experiments.

We next revisited a widely used DNA model, a neutral wormlike chain model, through a comparison of previous and new experimental data sets with the theory, as well as a dimensional analysis to provide a possible cause for the disagreement. With a set of *E. coli* data from Reinhart *et al.* [1] and a new set of λ-DNA data from genome mapping approach, we compared DNA extension distribution in experiments with theory. A breakdown was shown in the model as the channel size drops near or even below the persistence length of DNA in strong confinement. The discrepancy increases as the channel size decreases. Moreover, the treatment of using the alignment fluctuations or the effective channel size as fitting parameters fails to resolve it. Dimensional analysis of the wormlike chain propagator in channel confinement reveals the importance of a dimensionless parameter, reflecting the magnitude of the DNA-wall electrostatic interactions relative to thermal energy, that has not been considered explicitly in the prevailing theories for DNA confinement in a nanochannel. We thus propose that DNA-wall electrostatic interactions are the cause for the disagreement between experiments and theory and it has to be taken into consideration in a DNA model development.

We hope all the findings in this dissertation provide a deeper understanding of DNA properties and behavior in confinement and inspire scientists toward a new research direction to make a more robust DNA model for data interpretation of DNA-based experiments.

Contents

List of Tables

List of Figures

Statement of the Author's Contributions

Much of the research contained in this dissertation was originated from collaborative efforts. Some chapters have also appeared as articles in various journals. Because the articles list multiple authors, I will point out my contributions to the research.

I wrote Chapters 1 and 2. All the figures in Chapters 1 and 2 are adopted from publications without my contribution, and the related publications are cited in the corresponding figure captions.

Chapter 3 is based on H.-M. Chuang, J. G. Reifenberger, H. Cao, and K. D. Dorfman, "Sequence-Dependent Persistence Length of Long DNA", *Phys. Rev. Lett.* **119**, 227802 (2017). In this paper, we corrected an oversimplified assumption about how the persistence length of long DNA depends on the sequence and developed a statistical terpolymer model to rationalize the experimental results. In the approach, the data collection was done by J. G. Reifenberger at Bionano Genomics. I performed all data analysis and the model development. All the authors shared interpretation of the results. J. G. Reifenberger wrote the data collection part of the paper. K. D. Dorfman and I wrote the rest of the paper.

Chapter 4 is based on H.-M. Chuang, J. G. Reifenberger, A. B. Bhandari, and K. D. Dorfman, "Extension distribution for DNA confined in a nanochannel near the Odijk regime", *J. Chem. Phys.* **151**, 114903 (2019). In this paper, we reviewed a widely used DNA model by comparing the theory with experiments of two data sets. A possible cause for the disagreement between experiments and theory was proposed after a dimensional analysis was done. Both of the *E. coli* data set from Reinhart *et al.* [1] and the new λ-DNA data were collected by J. G. Reifenberger at Bionano Genomics. Some parameters used for the telegraph model in the paper were obtained by interpolation to the simulation data of Werner *et al.* [4] Three statistical tests were done by A. B. Bhandari. I did the rest of data analysis including data filtering, sensitivity analysis for λ-DNA data, and data comparison between two experiments and theory. K. D. Dorfman did the dimensional analysis. All the authors shared interpretation of the results. K. D. Dorfman and I wrote most of the paper.

I wrote Chapter 5. The data shown Fig. 5.2 was collected by J. G. Reifenberger, A. B. Bhandari and me at Bionano Genomics. I did the data analysis of Fig. 5.2.

Chapter 1

Introduction and Motivation

1.1 DNA Sequencing

Deoxyribonucleic acid (DNA) is the carrier of the genetic information of organisms. It has been studied intensely since 1953, when its three-dimensional double helical structure was found [5]. DNA is a natural biopolymer which is made up of repeating units called nucleotides. Each nucleotide is composed of a phosphate group, a sugar group and one of four nitrogen bases: adenine (A), thymine (T), guanine (G), and cytosine (C), as shown in Fig. 1.1. By reading the sequence of DNA, that is the order of base pairs (bp), genomic data can be obtained and decoded to translate into genetic information. Much effort has been devoted to studying DNA to understand the growth, development, functioning, reproduction, and evolution of living organisms [5,6]. The Human Genome Project (HGP) was launched in 1990 and declared completed in 2003 [7]. It was an international research project to sequence the human genome which involved a small number of individuals and assembled together results deducing the complete sequence for each chromosome as a reference. The assembled sequences show that extremely small differences in genotypes can lead to a large variation in phenotypes [8, 9]. This striking discovery motivated scientists to be further dedicated to mapping various genomes for genetic research.

The first sequencing technologies can be traced back to the 1970s with two-dimensional chromatography [12]. Shortly thereafter, different sequencing techniques, such as Maxam-Gilbert sequencing and Sanger sequencing, emerged to improve the accuracy and reduce

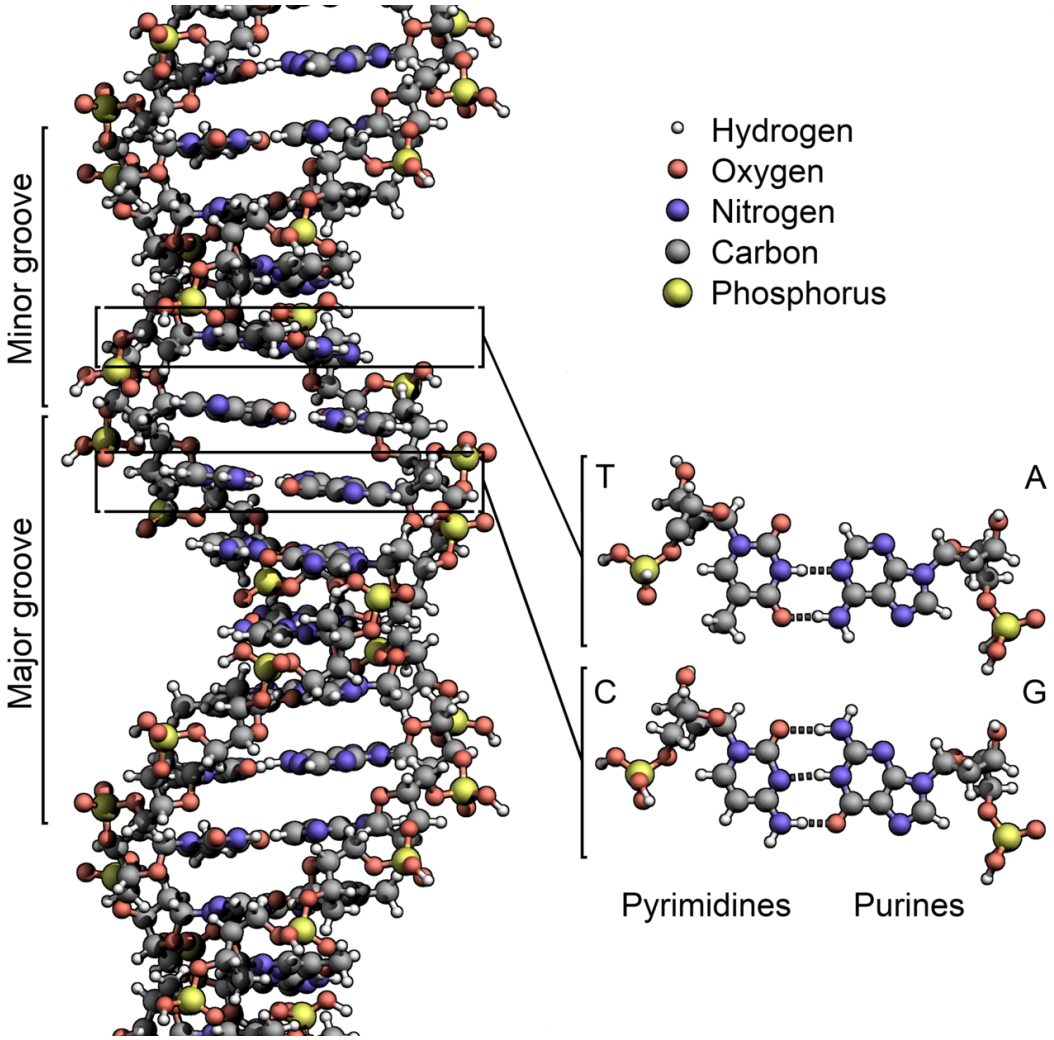

Figure 1.1: The chemical structure of the DNA double helix. The nitrogen bases of two strands are bound via hydrogen bonds according to base pairing rules: A with T, and G with C. Reproduced from wikipedia website [10].

the cost [13, 14]. Maxam-Gilbert sequencing, known as one of the first generation sequencing methods, was popular in the beginning but fell out of favor due to the radioactive labeling methods and toxic chemicals needed in the process [15]. Sanger sequencing, the other first generation sequencing method, was the major technique used in the HGP, and it can sequence DNA fragments up to 900 bp long at an error rate of 1% [16, 17]. Despite the merit of lower error rate and the reduction of hazardous chemicals needed

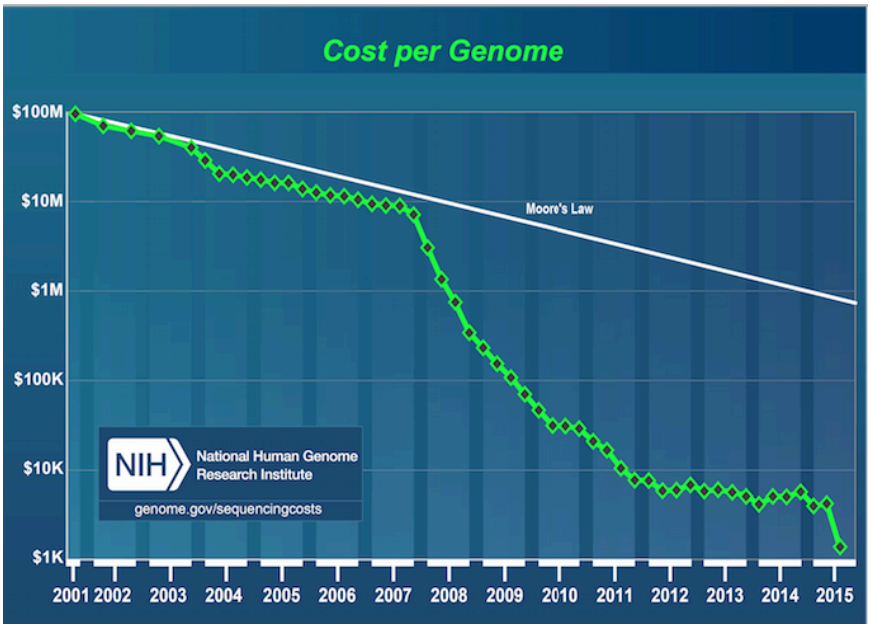

Figure 1.2: The cost needed to analyze a human size genome from 2001 to 2015. Reproduced from National Institutes of Health (NIH) website [11].

compared to the previous sequencing methods [13–15, 18], Sanger sequencing is a costly and time-consuming process and is difficult to scale-up. With the goal of improving these drawbacks of the Sanger sequencing techniques, Next-Generation Sequencing (NGS) was developed and became a popular tool to sequence DNA.

One of the main difference between Sanger sequencing and NGS is the sequencing volume. While Sanger sequencing can only sequence a single fragment at a time, NGS has a parallel system that sequences massive amounts of fragments simultaneously per run [20]. This benefit reduces the time and the cost needed to sequence a huge genome effectively. Figure 1.2 shows the cost needed to sequence a human size genome (3000 megabase pairs, Mbp) has reduced from $100,000,000 to about $1000 over the last 15 years [11]. This dramatic drop in cost, attributed to the technique of NGS [14, 21, 22], means that it is no longer a dream for the public to know their own DNA sequence at an affordable price. Despite the lower cost and the higher throughput, some challenges of NGS still exist and limit its development for further analysis. For example, NGS requires PCR to make copies of DNA segments, which will cause mutation and

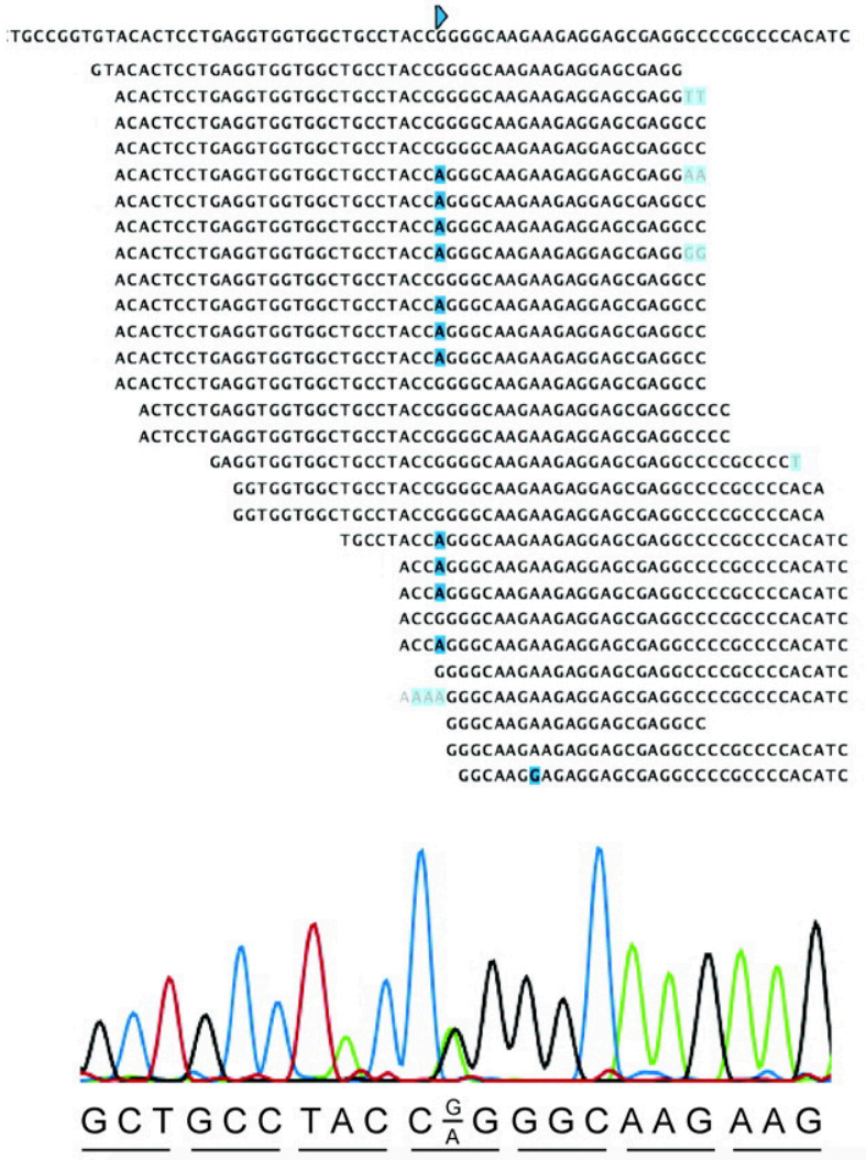

Figure 1.3: An example of how NGS aligns the sequences of 50 bp read length against the reference to detect a base pair mutation. Reproduced from Ref. [19].

amplification bias in product yield [23]. The read length of NGS has a limit of ∼300 bp, which means extremely high coverage is required to sequence a genome of large size as shown in Fig. 1.3, and simultaneously, it is insensitive to the structural variations of chromosome [24, 25]. Structural variation is the genomic alteration of an organism's

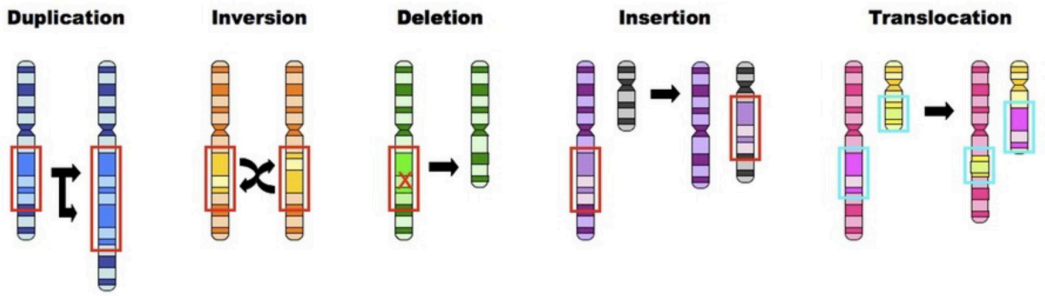

Figure 1.4: Overview of the different types of structural variations, including duplication, inversion, deletion, insertion, and translocation. Reproduced from BioNinja website [26].

chromosome. Common structural variations including duplication, inversion, deletion, insertion, and translocation, as shown in Fig. 1.4 [26], may occur on a sequence length between 1 kilobase pair (kbp) and 3 Mbp [27]. If we want to acquire information about structural variations by NGS, huge data storage space with required hardware is a challenge that needs to be overcome. Another powerful technique for sequencing DNA, optical mapping, was brought to the research field to detect structural variations at a large scale based on the need.

1.2 Optical Mapping

Optical mapping, first developed by Schwartz *et al.* [28], is a technique that images chromosome and DNA at a single molecule level using restriction enzymes. In this approach, large DNA fragments are elongated and immobilized on glass or molten agarose, treated with enzymatic digestion and cut into fragments. The size of DNA fragments can be measured through its fluorescence intensity profiles, and the corresponding molecular mass can be converted accordingly. These DNA pieces are collectively mapped as a unique 'fingerprint' or 'barcode' for that sequence by overlapping repeated fragments. The original goal of the technique is to preserve the internal order of a molecule so that the long-range information of a sequence can be obtained without amplification. When the technique combined with NGS, the resulting restriction map provides scaffolds for whole genome mapping and helps align the sequence in an order at a large scale, which eliminates the shortcoming of NGS. However, this method is limited by imprecise DNA

length measurement due to the non-uniform stretching [29], and the loss of information at the ends of DNA segments by the fragmentation [30]. Thus, DNA labeling barcode technique was developed by replacing restriction enzymes with nicking enzymes to generate barcodes [29]. DNA labeling barcode technique, unlike the restriction enzyme which fragments DNA, preserves the intact molecule by inserting fluorophores at sequence-specific nicks using nicking enzymes. After the labeling process, the backbone of DNA is stained with a fluorescent dye and stretched in a microfluidic device. In combination with fluorescence microscopy, the DNA backbone is detected and the distance between two adjacent labels can be measured. By knowing the distance between labels, and repeating the measurement with multiple overlapping long DNA fragments, it is possible to generate maps of parts of the genomes for matching against known references [29, 31–34]. Figure 1.5 shows the standard process to generate a labeling barcode for genome mapping. This map is critical to detect structural variations at a large scale as shown in Fig. 1.6. Structural variations at a length scale of kbp can be detected in a relatively short time, which NGS cannot achieve by its short read length.

While optical mapping resolves the insensitivity of NGS to chromosome structural variations, DNA linearization is another technical problem that shows up and remains challenging to scientists. Elongation of long DNA fragments is the key step of genome mapping. DNA is not rigid but a semiflexible polymer that has a randomized configuration due to thermal fluctuation. It coils and forms hairpins in free solution, but can be linearized with an external force or in confinement. Proper linearization of DNA helps reduce thermal fluctuation and thus the ordered information of optical maps can be read. In fact, through stretching DNA molecules, the resolution of mapping can be reduced from $\sim$10 Mbp for DNA in metaphase chromosomes to $\sim$1 kbp [35].

As mentioned earlier, restriction mapping needs DNA fragments to be elongated on the glass to read-out the order. Explicitly, this traditional 'combing'-based optical mapping uses fluid flows to elongate and fix molecules onto the surfaces of a pretreated glass [29, 36, 37]. In the original version of the process, the glass is first pretreated by having a layer of a vinyl silane on it. When the glass slide is dipped into the DNA solution, DNA molecules will attach to the glass surface by one end. As the slide is withdrawn, the DNA molecules can be stretched and aligned by a receding air-water interface and left to dry over the surface [37]. This process has been commercialized

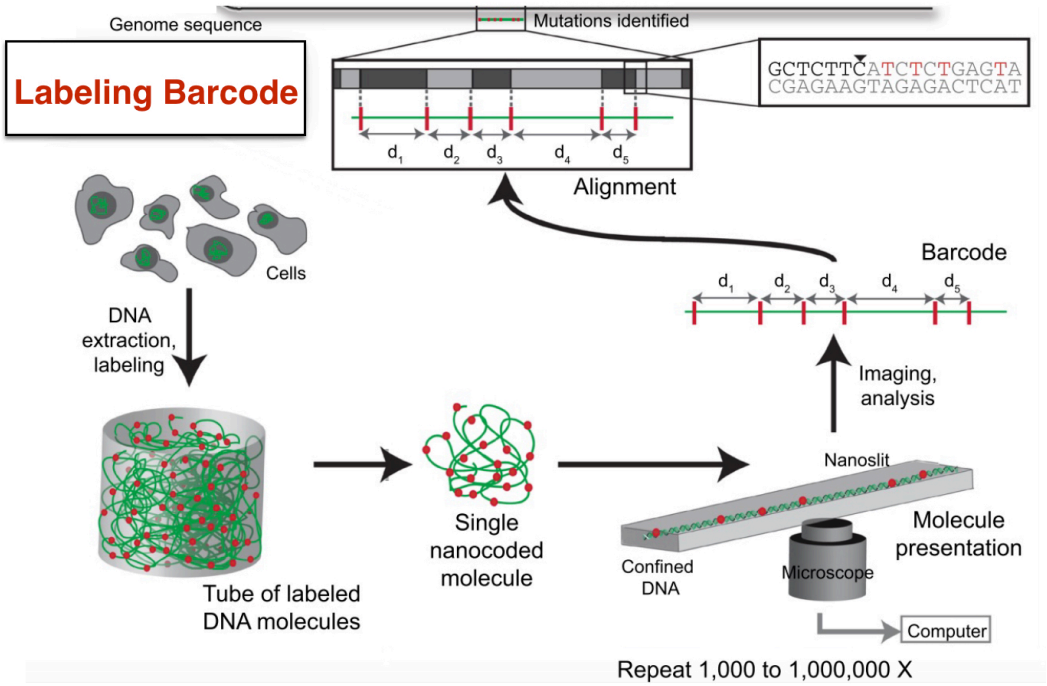

Figure 1.5: Long genomic DNA is extracted from the sample and nicked at sequence-specific sites with a nicking enzyme, such as Nb.BbvCI. The nicks are labeled with fluorophores and repaired by DNA ligase. Labeled DNA molecules are elongated in confinement and imaged using fluorescence microscopy to detect the labels on DNA backbone and a labeling barcode can be generated. The map can then be compared to a known reference and used as a scaffold for genome assembly and characterization. Reproduced from Ref. [30].

by companies Genomic Vision and OpGen. Another modified process with the similar idea develops fluid flows within a tiny, evaporating droplet to stretch and immobilize molecules onto the derivatized glass surfaces, as shown in Fig. 1.7 [36]. Although such evaporation-driven flow can produce extended molecules in general direction of the flow, stretched molecules often overlap due to random positioning (Fig. 1.7D), which makes the measurement and analysis extremely difficult especially for genomic DNA of large size [38].

Another approach of using external force to linearize DNA is based on a hydrodynamic focusing technique in microfluidics [29, 39]. In this method, double-stranded DNA is first tagged at sequence-specific motif sites with fluorescent bisPNA (Peptide

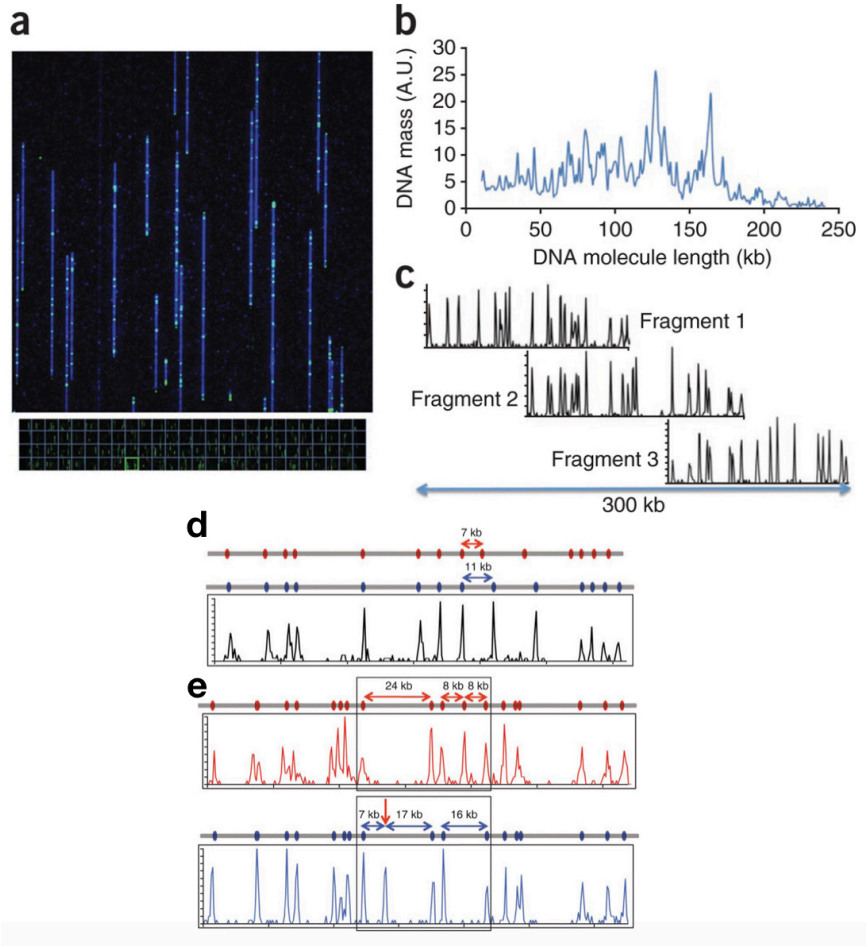

Figure 1.6: An example of structural variation detection by optical mapping technique. (a) Image of a single field of view containing a mixture of nick-labeled DNA molecules in the nanoarray. (b) The distribution of the DNA molecules imaged on the nanoarray by length. The majority of the molecules are 100-170 kbp in length. (c) After clustering of DNA molecules based on nick-labeling patterns, consensus maps with overlapping patterns are assembled into contiguous-sequence motif maps. (d) The Nt.BspQI map indicates that the sample genome (gray line with red dots) has a 4 kbp deletion as compared with the reference genome (gray line with blue dots), with a 7 kbp and an 11 kbp fragment between two neighboring sites. (e) An Nt.BspQI site identified in the region (arrow) is found in the reference genome (gray line with blue dots), splitting the 24 kbp fragment into 7 kbp and 17 kbp fragments. The sample genome produced by genome mapping (gray line with red dots) shows a 24 kbp fragment with a haplotype variation in the adjacent region. Reproduced from Ref. [29].

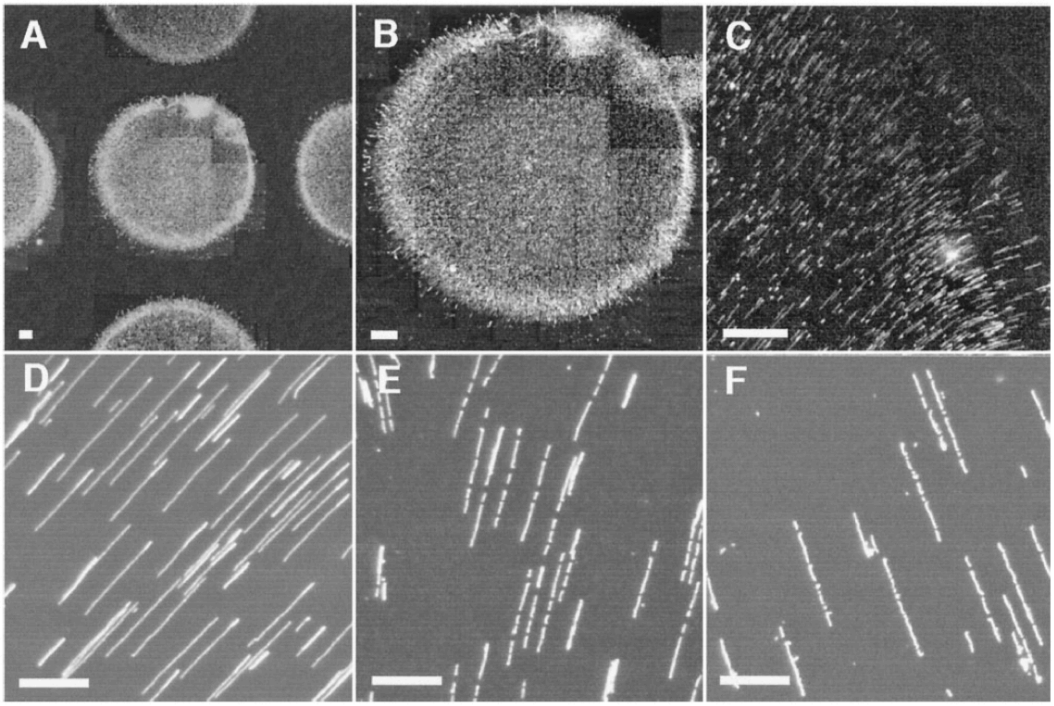

Figure 1.7: λ-DNA was dissolved in Tris-EDTA buffer containing 0.5% glycerol deposited onto APTES-treated glass surfaces, dried and stained. (A) An image of droplets on a derivatized surface. (B) Close-up of a droplet within the grid. (C) Elongated DNA molecules on the surface before restriction digestion. (D) Magnified image of elongated DNA molecules contained within the droplet shown in B before restriction digestion. (E) DNA molecules in B, different field, after digestion with a restriction enzyme. Note the appearance of gaps signaling enzyme cleavage sites. (F) DNA molecules after digestion with another restriction enzyme. [Bars: 20 μm (A-C); 5 μm (D-F).] Reproduced from Ref. [36].

Nucleic Acid). The labeled DNA is then stretched in a single channel microscale device by elongation flow and the sequence-specific fluorescent tags are detected by a multicolor detection system [39]. Although this technique, also known as Direct Linear Analysis (DLA), elongates and aligns longer DNA molecules without overlap [40], the non-uniform stretching and limited throughput still remain challenges that need to be conquered for large-scale genome mapping with accuracy [41, 42].

In addition to external force, loading DNA in confinement is an alternative to linearizing DNA. Figure 1.9 shows how a coil structure of a long DNA molecule can be

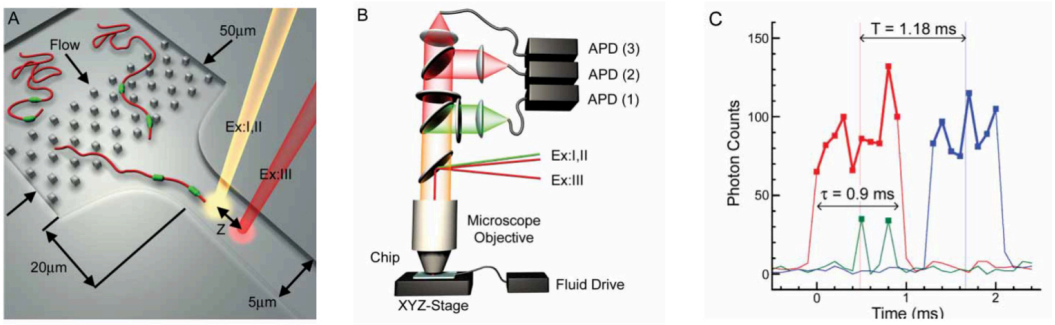

Figure 1.8: Schematics of the Direct Linear Analysis technology. (A) A cross-section of the microfluidic device (top view). (B) Optical scheme (side view). The excitation and detection are arranged within a confocal fluorescence microscope. (C) Typical raw data traces from channels 1-3 for a single tagged DNA molecule. The red and blue traces arise from the fluorescence of the intercalating dye when the DNA backbone travels through the excitation spots ExII and ExIII, respectively. The green spikes are detected when the DNA-bound PNA tags pass through the excitation spot ExI and emit bursts of photons. Reproduced from Ref. [39].

linearized as the scale of confinement drops [43]. High-throughput and high resolution of the map are the goals when we develop an optical genome mapping technique. A nanochannel array used as a platform to generate labeling barcodes can provide high-throughput results and avoid the drawbacks of the traditional methods to elongate DNA. There are several types of geometries a nanochannel array could take, such as nanoslit, nanotube, triangular nanochannel, rectangle or square nanochannel, and spatial gradient structures. Several studies have shown that not only the configuration but also the physical properties of DNA are greatly affected in such confinement [44, 45]. In order to correctly interpret experimental results from optical mapping using nanochannels for further clinical and genetic research, deep understanding of static and dynamic properties of DNA in confinement is required.

In combination with proper confinement, optical mapping becomes a powerful tool to generate high-throughput labeling barcodes for large-scale genome mapping. For example, the company Bionano Genomics commercializes the technique as a series of products such as nanochannel array, DNA reagents for experiments, a laser system paired with a fluorescence detector to detect fluorescence labels, and corresponding bioinformatic

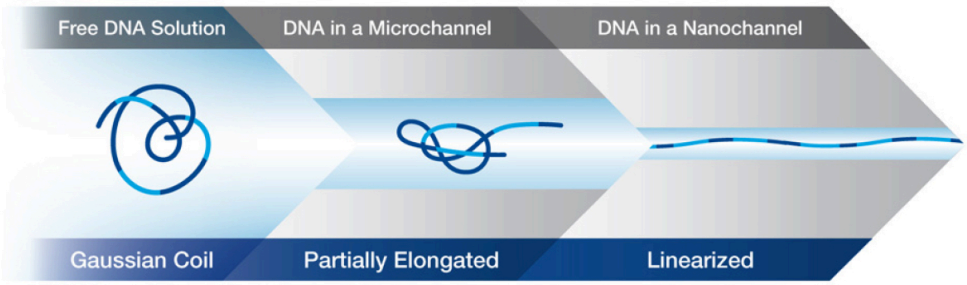

Figure 1.9: The coil structure of a long DNA molecule in free solution can be linearized through nanoscale confinement. Reproduced from Bionano Genomics website [43].

software for consumers to analyze targeted DNA by themselves [43]. The result obtained from the technique provides good scaffolds for structural variation detection and whole genome mapping when coupled by NGS [46, 47].

1.3 Research Outline

Nowadays, after decades of improvement, optical mapping with DNA linearization in confinement has been a well-developed technique which has been widely used for scientific research in multiple fields. The next step to make it even better is focused on the data interpretation. It is worthwhile to answer some questions like "How DNA extends in confinement?" and "How thermal fluctuations are reduced in that confinement?" before further analysis. As mentioned in the last section, genetic information needs to be translated from these experimental results through correct data interpretation. How the data are analyzed relies on the comprehensive understanding of DNA behavior in confinement and the corresponding physical properties, and all these are like a huge puzzle where each finding is a piece of it. In fact, physicists have already put a lot of effort to try to find every piece of the puzzle. For example, four regimes as a function of channel size were successively proposed to give possible predictions of how DNA behaves in certain confinement [44, 48–51]. In addition, the change in the physical properties of DNA with experimental condition, like how the rigidity of DNA is affected by the salt concentration of buffer solution, was also well-studied [52,53]. However, even with these pieces of the puzzle at hand, there is still some disagreement between experiment and

theory which cannot be fully explained. *As such, the goal of this dissertation research focuses on finding the missing pieces of the puzzle. In particular, we aim to correct an oversimplified assumption about the properties of DNA, review a widely used model which describes the behavior of DNA in confinement with an analysis of using a previous and new sets of experimental data, and propose a possible cause for the disagreement between experiment and theory.* Moreover, conventional methods which were used to study DNA properties at a single DNA level, such as neutron scattering [54], magnetic tweezers [55] and atomic force microscopy (AFM) [56], are usually too laborious and time-consuming to collect enough data points for making any conclusion. We thus adopted the high-throughput labeling barcode technique developed by Bionano Genomics for the data collection in this whole research work. Once the high-throughput data were obtained, relevant statistical analyses were applied to interpret the results. By combining DNA physics and statistics, two disciplines which did not cross over in the past, a new insight about the physical properties of DNA was offered with trustworthy data.

In line with the above goals, **Chapter 2** provides a brief background and reviews recent literature concerning the physics of confined DNA, which are necessary to understand the materials presented in the following chapters. We begin with the introduction of several key physical parameters of DNA in confinement and summarizes how these properties depend on the sequence and the ionic strength of the solution. We then provide an overview of all stages of nanochannel confinement, and focus on the two regimes which are relevant to the dissertation research.

With the basic background summarized in the last chapter, **Chapter 3** is the first missing piece of the puzzle found by correcting an oversimplified assumption about one of the DNA properties, the persistence length, in the model we usually use for long DNA. People usually assume there is no dependence of the persistence length on DNA sequence when DNA is long enough to average over the intrinsic curvature. We correct this rough assumption through the approach of data collection, result interpretation, and model development to predict the persistence length of any long DNA sequence. Explicitly, we explain how we worked collaboratively with Bionano Genomics to gain high-throughput data and employed relevant statistical analyses to make a convincing conclusion. We subsequently develop a theoretical statistical terpolymer model which consists of an intrinsic sequence-dependent term and an electrostatic sequence-independent term to

rationalize the strong correlation between the persistence length and DNA sequence found in the experiment. This model helps estimate the persistence length of arbitrary DNA sequence with known base pair composition and experimental condition. We are optimistic that the model proposed will prove useful for quantitative analysis of DNA-based experiments.

The next missing piece of the puzzle is found by reviewing a robust DNA model, a neutral wormlike chain model, through an analysis of using previous and new sets of experimental data. In **Chapter 4**, we compare the data with the theory and provide a possible cause for the discrepancy. DNA confinement in a nanochannel typically is understood via mapping to the confinement of an equivalent neutral polymer by hard walls. This model has proven to be effective for confinement in relatively large channels where hairpin formation is frequent. An analysis of existing experimental data for *E. coli* DNA extension in channels smaller than the persistence length, combined with an additional data set for λ-DNA confined in a 34 nm wide channel obtained through a high-throughput genome mapping technique, reveals a breakdown in this approach as the channel size approaches the Odijk regime of strong confinement. In particular, the predicted extension distribution obtained from the asymptotic solution to the weakly-correlated telegraph model for a confined wormlike chain deviates significantly from the experimental distribution obtained for DNA confinement in the 34 nm channel, and the discrepancy cannot be resolved by treating the alignment fluctuations or the effective channel size as fitting parameters. We posit that the DNA-wall electrostatic interactions, which plays an important role in governing the extension of DNA for channels close to the persistence length, are the source of the disagreement between theory and experiment by providing a dimensional analysis to reinforce our argument.

Finally, **Chapter 5** summarizes the important discoveries of this dissertation and proposes possible works for future research directions with some preliminary data and corresponding analysis. In particular, this chapter highlights the gaps in the puzzle which have not been solved yet but are the most relevant to this work.

Chapter 2

Background

This chapter provides a brief background and reviews recent literature concerning the physics of confined DNA, which are necessary to understand the materials presented in the following chapters. Section 2.1 begins with the introduction of several key physical parameters of DNA in confinement and summarizes how these properties depend on the sequence and the ionic strength of the solution. Section 2.2 provides an overview of all stages of nanochannel confinement, and focuses on the two regimes which are relevant to the dissertation research. For more details, please see the review article by Reisner *et al.* [44]

2.1 Fundamental Physical Properties of DNA

A polymer is a macromolecule composed of numerous repeating monomers. Different chemistries and various amounts of repeating monomers endow polymers with unique physical properties. DNA is a natural biopolymer whose repeating unit is nucleotide. Over the past decades, physicists have made concerted efforts to understand the physical properties of polymers by using DNA as an important model system. Long double-stranded DNA (dsDNA) molecules can be visualized and studied at a single molecule level by staining the molecule and using fluorescence microscopy [57, 58]. To properly study and interpret the experimental results, several key physical parameters characterizing DNA were defined: the contour length (L), the persistence length (l_{p}), the DNA effective width (w) and the wall-DNA depletion width (δ). Figure 2.1 shows a schematic

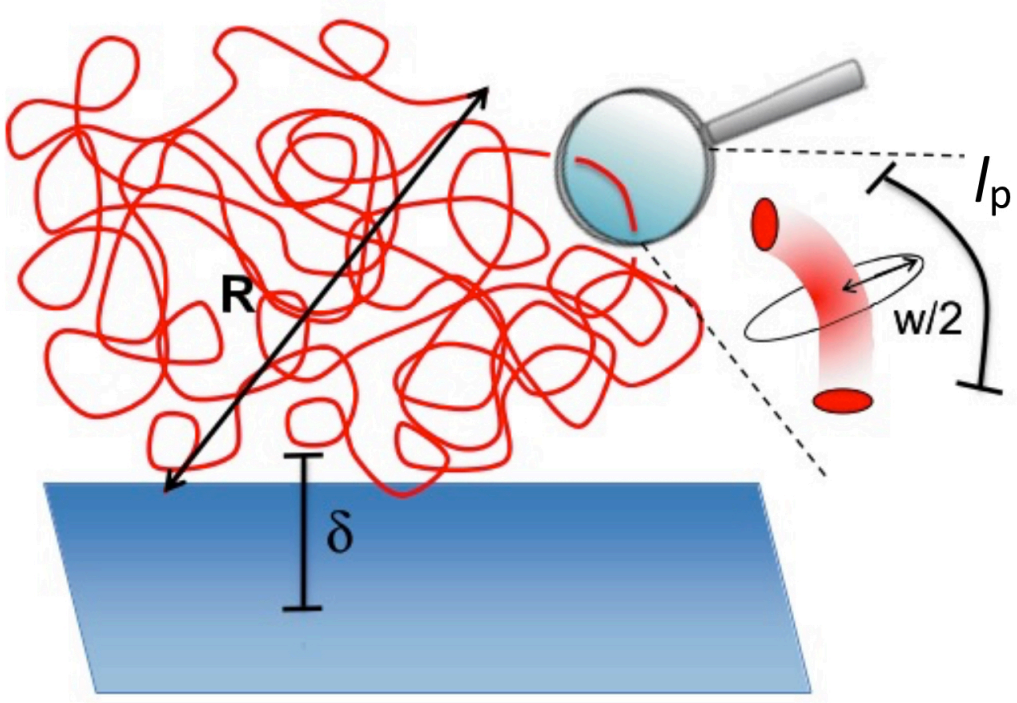

Figure 2.1: Schematic of DNA as a wormlike chain with the end-to-end length (R), persistence length (l_{p}), effective width (w) and wall-DNA depletion width (δ). Reproduced from Ref. [44].

of a typical DNA conformation with most parameters labeled. Below we provide the introduction to each parameter.

2.1.1 Contour length (L)

The contour length is the length of end-to-end distance, R, of a DNA molecule stretched without thermal fluctuation. It can be calculated simply by taking the total number of base pairs in a molecule and multiplying by the average rise of a base pair. For B-state dsDNA, the average rise of a base pair is 0.34 nm [59]. Let's take λ-DNA as an example, which is the most commonly used DNA for DNA experiments. Unstained λ-DNA with 48,510 bp in total, has a contour length of 16.5 μm. When doing genome mapping and single DNA analysis, in order to make DNA visible with fluorescence microscopy, fluorescent dyes like YOYO-1 or TOTO-1 are used to stain molecules to

enhance signal by intercalating binding. The bis-intercalation of fluorescent dyes is well-known to increase the contour length of DNA [60,61], which is an issue we need to take into consideration when we interpret the results. In this dissertation research, we used low dye loading (1 dye molecule per 37 bp) in experiments to reduce the effect caused by the binding dye on the contour length.

2.1.2 Persistence length (l_{p})

DNA is not rigid but a semiflexible polymer that has a randomized configuration due to thermal fluctuation. The persistence length, which can capture the rigidity of DNA, has multiple definitions based on the mathematical description. One is connected to the molecular architecture of the atoms, which maps the persistence length to the thermal energy ($k_{\mathrm{B}}T$, where k_{B} is the Boltzmann constant and T is the absolute temperature) and the dimension of space at the atomic scale [44]. Another commonly used definition is defined by the distance over which two segments of the chain remain directionally correlated [62,63]. There are several experimental methods to measure the persistence length of DNA, including light or neutron scattering (see reference in [64]), magnetic tweezers [55], and AFM [56]. All of these methods return a value of persistence length by accounting for a balance between energy and entropy [44].

Persistence length is significantly dependent on the salt concentration of the solvent. The salt concentration, which can be converted into ionic strength, can cause up to a 60% difference to DNA persistence length between high ($>$ 100 mM, l_{p} = 50 nm) and low ($<$ 10 mM, l_{p} = 80 nm) ionic strength buffer [52,53].

Classical theory suggests that the measured persistence length is the sum of the bare persistence length and the electrostatic contribution, which can be written as the following formula [52,53,65]

$$l_{\mathrm{p}} = l_{\mathrm{p,0}} + A\lambda_{\mathrm{D}}^2, \tag{2.1}$$

where $l_{\mathrm{p,0}}$ represents a 'bare' persistence length which can be estimated at a high salt buffer solution. A is a prefactor that is simplified by Manning condensation theory to account for charge density and Bjerrum length [66]. λ_{D} is Debye length that is affected by the ionic strength. Though Eq. (2.1) suggests $l_{\mathrm{p}} \sim \lambda_{\mathrm{D}}^2$, Dobrynin posited $l_{\mathrm{p}} \sim \lambda_{\mathrm{D}}$ instead by arguing the deformation mode which overestimated the electrostatic energy

penalty due to chain bending in Eq. (2.1) [67]. By fitting to experimental data for λ-DNA, Dobrynin obtained an empirical formula

$$l_{\mathrm{p}} \approx 46.1\mathrm{nm} + 6.3\lambda_{\mathrm{D}}. \tag{2.2}$$

It is hard to validate which definition is the correct one based on the existing data. In the following chapters, we chose to apply the Eq. (2.2) from Dobrynin and the calculated ionic strength of our experimental results to estimate the electrostatic contribution of solution to l_{p}.

It is reasonable to question if the persistence length of DNA varies with base pair sequence, and the answer is absolutely yes. The bare persistence length, $l_{\mathrm{p},0}$, accounts for the elastic moduli of the sequence itself. A previous study has shown that elastic constants of the DNA helix are associated with DNA sequence [68]. Explicitly, four-fold difference in the elastic constants of DNA helix was found for the sequences of dG-dC compared to the sequence of random dAT tracts [68]. Biophysicists also have found that some specific sequences of around 100 bp have unique intrinsic curvature which is critical to nucleosome positioning [69–72]. The dependence of intrinsic curvature on sequence implies that the bending energies between nucleotides differ substantially. This difference should manifest at long length scales in the DNA persistence length in the same way that hindered rotation around carbon-carbon bonds leads to a 13% increase in statistical segment length for polystyrene when compared to polyethylene [73]. For example, while a random sequence has the persistence length of 64 nm long, the relatively large persistence length of 84 nm was found for poly(dG − dC)·poly(dG − dC) under the same experimental condition [74]. Though the persistence length of DNA must depend on the sequence, physical experiments using long DNA (>2000 bp) typically assume it is a weak function of sequences for simplicity. The potential inaccuracy of data analysis caused by using overly simplistic models for the DNA persistence length motivated this research study. To understand how the persistence length of long DNA depends on the sequence composition, we conducted experiments using a high-throughput genome mapping technique with human DNA and proposed a statistical terpolymer model to rationalize our experimental results in Chapter 3.

2.1.3 Effective width (w)

The effective width quantifies the repulsive interaction between negatively charged DNA segments, as shown in Fig. 2.1. The screening of backbone charges by the counterions in the solution will affect not only the l_{p} but also the w of DNA. Both l_{p} and w increase when the ionic strength of the buffer decreases, which makes the DNA stiffer. [64] The standard approximation to predict the w was proposed by Stigter [75] using Poisson-Boltzmann theory to calculate a rod-rod interaction potential, and then evaluating the rod-rod excluded volume by the interaction. The estimation agrees well with the experimental results using techniques such as light scattering, sedimentation and measurements of the probability of DNA knotting during cyclization [76–78]. Both experiments and theory suggest that the effective width is highly dependent on the ionic strength, varying from 5 nm at high concentration (100 mM) to 20 nm at low ionic strength ($<$ 10 mM) [44].

2.1.4 Wall-DNA depletion width (δ)

The wall-DNA depletion width is used to explain the repulsions between DNA and the device walls, which includes the hardcore interaction and the electrostatic contribution if the wall is also negatively charged [79], as shown in Fig. 2.1. Knowing the depletion width is important since the effective channel size, D_{eff}, which is the actual cross-sectional width where DNA is accessible, is defined as

$$D_{\mathrm{eff}} = D - \delta, \tag{2.3}$$

where D is the measured channel size. Many studies equate the depletion length with the effective width to approximate D_{eff} by modeling the DNA-wall repulsions as intersegmental DNA interactions. This approximation yields around 10% error between simulation and experiments [80], and this uncertainty is usually explained away as a source of the systematic error when interpreting the results [44, 81, 82]. In addition to the approximation of $\delta = w$, there are two other models for the estimation of the depletion length. In the model proposed by Derek Stein [44], DNA is considered as a charged, semiflexible polymer interacting with the electric potential created by the wall. The other model, which was proposed by Reisner *et al.* [44], regards DNA as a charged,

rigid rod instead of a semiflexible chain with hindered rotation due to the wall. Among these three models, none of them can fully explain the results of every experiment. It is the complex dependence of the depletion width on the ionic strength that makes the complete resolution of the disagreement difficult.

Our group recently conducted a series of experiments using the Bionano Genomics Irys platform to study the dependence of the depletion width on the ionic strength [83]. We generated high-throughput data by a genome mapping technique [84–86] and analyzed the existing models for the depletion lengths with the measurements. A simple model for the depletion length of DNA in a high ionic strength buffer is built in the work [83] and was used to re-evaluate data from previous experiments. We found it surprising that the deviations between the theory and experiments are increased by approximately 20% after the approximation of D_{eff} was corrected by the calculated δ from our model. The increased deviation means that the rough approximation of $\delta = w$ is not the main factor for the disagreement between the theory and experiments. This result (Fig. 7 of Ref. [83]) motivates us to further investigate the potential sources of systematic error in experimental studies and mapping technology. There should be some other systematic error sources, such as the unknown dye loading effect on the persistence length of DNA, which were reported by conflicting results [87–92] in experimental studies, or the oversimplified wormlike chain model which ignores the local electrostatic interactions between DNA and the wall. The results reported by our group recently suggest re-examining the basic physical properties of DNA in a confinement. We thus reanalyzed the data by Reinhart *et al.* [1], conducted another set of experiments using the genome mapping technique with λ-DNA to remove all possible artifacts in previous experiments, and re-analyzed the theory of DNA confined in a nanochannel near the Odijk regime in Chapter 4. For more detailed information about the models of the depletion width, see Ref. [83].

2.2 DNA in Nanochannel Confinement

As mentioned in the last section, DNA is a popular model for researchers to study polymer physics. With proper confinement and the use of fluorescence microscopy, DNA can be visualized and its static as well as dynamic properties can be observed. DNA

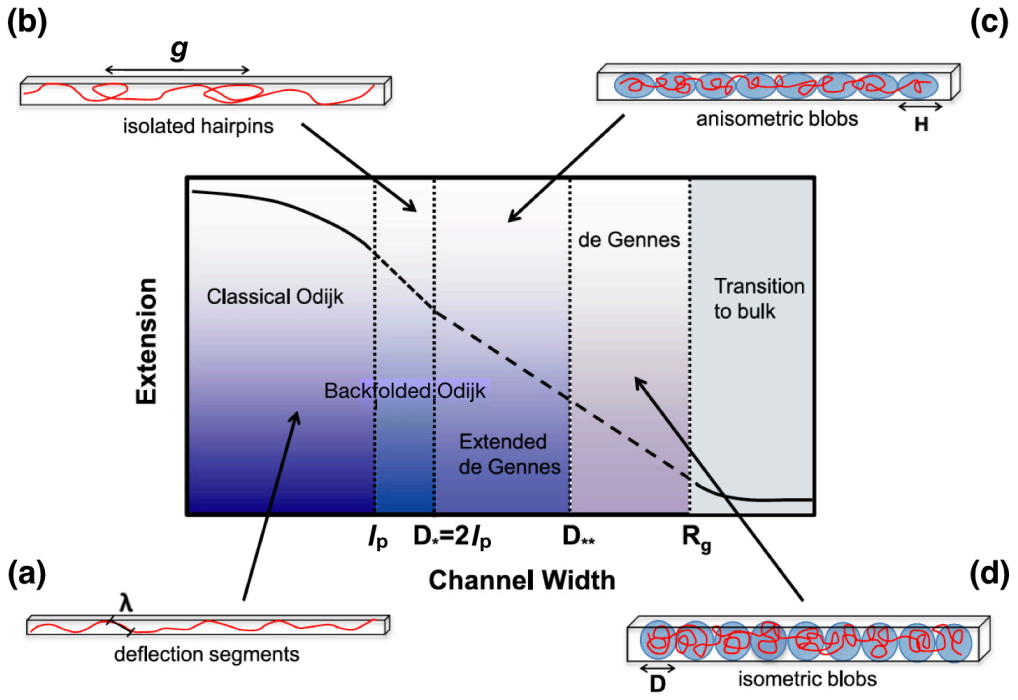

Figure 2.2: Schematic of four different regimes in nanochannel confinements with increasing channel widths D, (a) the Odijk regime, (b) the backfolded Odijk regime, (c) the extended de Gennes regime, and (d) the de Gennes regime. Reproduced from Ref. [44].

confinement in a nanochannel is typically understood via mapping to the confinement of an equivalent neutral polymer by hard walls. There are multiple regimes for a square nanochannel confinement in a width of D. Figure 2.2 shows four different regimes in nanochannel confinements with increasing channel widths. In Fig. 2.2(d) where $D \gg l_{\mathrm{p}}$, known as the de Gennes regime, the semiflexible wormlike chain is weakly affected by the confinement and the molecule coils like a series of blobs with the conformation in each blob is a self-avoiding random walk. In this regime, the radius assumes the channel width up to an upper limit of the radius of gyration, R_{g}, in solution [48]. The conformation of the polymer is dominated by the hard-core repulsion of the intersegmental DNA interactions with the effective width w, which is defined in the last section. As the width of the confinement decreases, the influence of the wall becomes stronger. In the lower limit where $D \ll l_{\mathrm{p}}$, also known as the Odijk regime (Fig. 2.2(a)), the bending

energy increases and the molecule can no longer coil or back fold to form a hairpin. In this regime, the chain can produce successive deflection segments with a special length scale λ called the 'Odijk deflection length' [93], and is extended close to its contour length, with small fluctuations in alignment with respect to the channel axis [49,94,95].

Within the two limiting cases, there are two regimes that have not been introduced yet, the backfolded Odijk regime, and the extended de Gennes regime, as shown in Fig. 2.2(b) and (c), respectively. As the channel width decreases below the critical length of the de Gennes regime, that is the D_{**} shown in the Fig. 2.2, Odijk argued that the isometric blobs in the de Gennes regime will become anisometric [49]. In the extended de Gennes regime, the excluded-volume interaction is still significant and dominates the linear ordering of blobs, but not strong enough to drive the blob statistics as the one in the de Gennes regime [44]. As the channel width keeps decreasing and drops below D_*, roughly equal to $2l_\mathrm{p}$, but still beyond the Odijk regime, the wall effect increases and the entropy reduces. In this regime, the space is still small enough that the polymer chain can produce deflection segments, but also large enough that it can form hairpins. Odijk [49] proposed the scaling behavior in this so-called backfolded Odijk regime requires that the polymer exhibits a very large monomer anisotropy [51], the latter measured by the ratio of the persistence length to the effective width of the polymer. For DNA, the monomer anisotropy l_p/w saturates at $l_\mathrm{p}/w \approx 10$ at high ionic strength and decreases with decreasing ionic strength [64]. As a result, DNA will not exhibit a backfolded Odijk regime [51]. Rather, the extension of DNA confined in channels near the persistence length typically appears as a smooth transition from the de Gennes scaling to the Odijk scaling, with an apparent extension scaling that is inverse of the channel size [96–101].

Recently, two complementary theories have emerged that subsume the physics of the extended de Gennes regime and the backfolded Odijk regime into a single regime characterized by weak excluded volume interactions [4, 102]. The key result of both theories [4, 102] is the existence of a new scaling parameter, α, that represents the typical number of overlaps per hairpin bend in the chain, with the fractional extension of the chain scaling like $\alpha^{1/3}$. Importantly, the scaling of the fractional extension is not predicated on a large value of the monomer anisotropy. As a result, these theories [4,102] should permit a description of DNA extension when $D \approx l_\mathrm{p}$, a technologically important

case [29] for which the inequalities required in Odijk scaling theories [49] cannot be satisfied.

The data discussed in this dissertation research were all measured in nanochannels ranging from 34 nm to 51 nm, which are somewhat smaller than the persistence length of DNA ($\approx$ 50 nm in a high ionic strength buffer), and thus being proximate to the Odijk regime but not yet satisfying the strong inequality $D \ll l_{\mathrm{p}}$. In principle, we could take advantage of Odijk scaling and one of the new theories, the weakly-correlated telegraph model [4], to facilitate the analysis of the distribution of extension and other physical properties of DNA for further study. However, a breakdown was revealed in the approach as the channel size approaches the Odijk regime of strong confinement in this dissertation research. We posit that the DNA-wall electrostatic interactions, which are sensible throughout a significant fraction of the channel cross-section in the Odijk regime, is the source of the disagreement between theory and experiment. For more detail about the weakly-correlated telegraph model, and the discussion on the disagreement between the theory and experimental studies for DNA confined in a nanochannel near the Odijk regime, see Chapter 4.

Chapter 3

Sequence-Dependent Persistence Length of Long DNA

This chapter is based on the publication
H.-M. Chuang, J. G. Reifenberger, H. Cao, and K. D. Dorfman, "Sequence-Dependent Persistence Length of Long DNA", *Phys. Rev. Lett.* **119**, 227802 (2017). [103]

3.1 Introduction

Over the past two decades, long molecules of double-stranded DNA have emerged as an important model system in polymer physics, with applications in rheology [57,104–106], confined polymers [38, 44, 79, 96, 107], and transport in model porous media [38, 108]. A particularly salient advantage of DNA is the ability to visualize the polymer by fluorescence microscopy, thereby directly interrogating the underlying physical models at the single-molecule level. The proper interpretation of these experiments requires an accurate measurement of the DNA persistence length. Often, the persistence length is obtained from force-extension experiments [55] or polyelectrolyte theory [67]. These approaches often assume that the persistence length of DNA is, at most, a weak function of sequence. In this chapter, we present data obtained from a high-throughput genomic mapping method [1, 29] that calls into question this widespread assumption. Using circa 5×10^7 measurements of DNA extension in nanochannels, we show that the 2% increase in fractional extension as % GC content increases (which does not affect the

genome mapping strategy employed here) translates into a persistence length that varies by almost 20% due to the relatively weak dependence of the fractional extension on persistence length in the Odijk regime [93]. Building on existing concepts [67, 68], we rationalize our results by modeling long DNA as a statistical terpolymer with a sequence-dependent intrinsic persistence length.

3.2 Background

The neglect of DNA sequence in many polymer physics experiments stands in stark contrast to that in biophysics. The so-called "intrinsic curvature" of DNA, which emerges over circa 100 base pairs, depends strongly on DNA sequence [2,68,109] and is purported to play a role in biological processes such as nucleosome positioning [69–72]. Likewise, certain sequences such as poly(A) tracts introduce local bends in DNA [56, 110–112], again at very short length scales. These local properties are modeled by a sequence-dependent bending energy that depends on the dinucleotide pair being bent [2,68]. The dependence of intrinsic curvature on sequence implies, *inter alia*, that the dinucleotide bending energies differ substantially. As such, they should manifest at long length scales in the DNA persistence length in the same way that hindered rotation around carbon-carbon bonds leads to a 13% increase in statistical segment length for polystyrene when compared to polyethyelene [73].

3.3 Data Collection

3.3.1 Nanochannel Experiments

Measuring how the DNA persistence length depends on sequence, while simultaneously ensuring the sequence is long enough to average over the intrinsic curvature, is an onerous task. Standard methods, such as light or neutron scattering (see references in [64]), magnetic tweezers [55] and AFM [56] are inherently low-throughput. We thus adopted the genome-mapping approach described in Fig. 3.1.

Briefly, DNA was prepared from a cell line created for the HAPMAP project, NA12878 from the CEPH Utah Reference collection. The sample's origin is from a Caucasian female. The DNA were nick-labeled using Nt·BspQI (New England Biolabs)

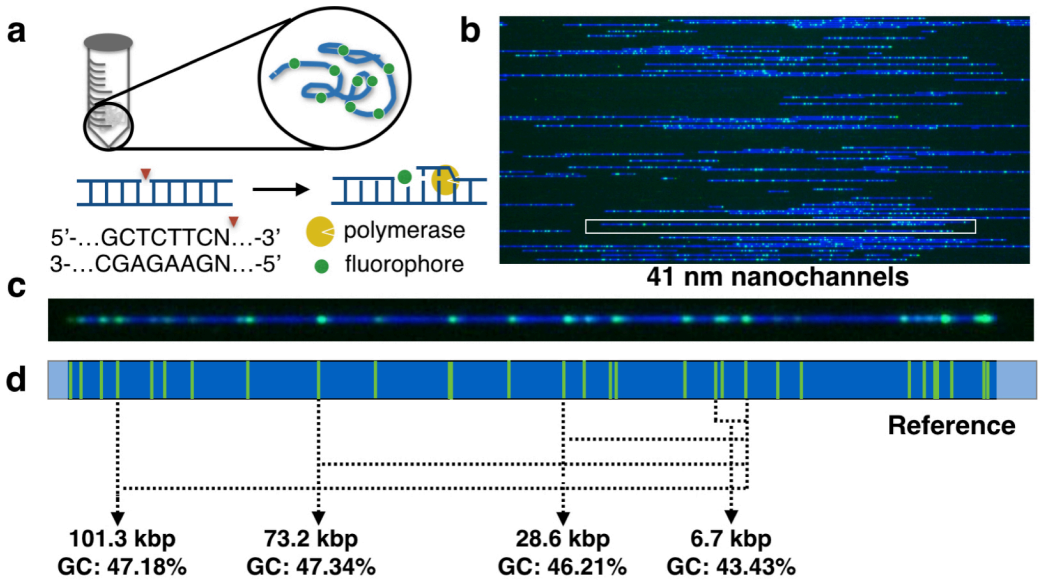

Figure 3.1: Experimental approach to measure the persistence length of DNA over a wide range of sequences. (a) Human DNA are fluorescently labeled at the GCTCTTC sequence by nick labeling and the backbone is stained with YOYO-1. (b) The labeled DNA are stretched in a 41 nm × 41 nm nanochannel using the high-throughput Irys genome mapping system. (c) Individual molecules are mapped to the (d) reference human genome, which reveals the underlying sequence – and % GC content – between nick sites. This particular molecule (154.6 kbp) has 30 nick sites; the % GC content values for 4 of the 435 possible pairs of nicking sites on this molecule are indicated.

to insert cy-3-like fluorescent nucleotides at the nick site GCTCTTC [29], and the backbone was stained with YOYO-1 (Invitrogen) at a ratio of 1 dye molecule to 37 base pairs (Fig. 3.1a) [86]. The DNA were then stretched by electrokinetic injection into an array of 41 nm wide, square nanochannels on an Irys v2 chip (BioNano Genomics) and imaged on a research-grade version of the Irys system (Fig. 3.1b) using the IrysPrep buffer (BioNano Genomics, ionic strength = 48 mM). We obtained data on 452,219 DNA molecules at least 150 kbp in size. Individual molecules were kept for further analysis if they met the following two criteria: 9 or more nick sites on the molecules and aligned to the reference with a p-value $\leq 10^{-9}$. The data aligned at a rate of 85% resulting in a coverage of 36×. We only considered the extension between pairs of nick sites in a given chromosome that are (i) separated by at least 2.5 kbp (for adequate resolution) and 393 kbp (for adequate sampling) and (ii) do not contain any N-base (unknown)

regions in the human genome. Removing N-base regions is essential, as these unknown sequences in the reference genome introduce systematic errors [113]. Figure 3.1c shows a representative molecule with 30 nick sites; the % GC content for the sequences between 4 of the 435 possible pairs of nicking sites on this molecule are indicated in Fig. 3.1d.

3.3.2 Selection of Nick Pairs

In the first pass through the data, we required that all nick sites from the human reference needed to be a resolvable site with no chance of resolution interference. Nearby nicking sites can appear as a single dot due to the diffraction limited optics in the system. In typical nanochannel mapping experiments, the ability to resolve pairs of nick sites is limited by the optics. The ability to distinguish the two fluorescent "dots" increases with increasing genomic distance between them until reaching a plateau at a separation of 2500 bp. Note that nearby labels (say, 3 kbp apart) that are not resolved are unlikely to have a significant effect on the analysis; the computed stretching the single dot and some other label in the 5' direction would be 500 bp shorter but a similar measurement between the single dot and some label in the 3' direction would be 500 bp longer (or vice versa). Such systematic errors would likely average out, and the effect certainly becomes diluted as the distance between this single dot and some other label becomes large.

We thus searched the human genome for singly resolvable nick sites between 2.5 kbp apart and 500 kbp apart and calculated the % GC content between each pair based on the reference genome. This search produced 5,563,589 possible pairs of nick sites. We then further refined our search by requiring that a given pair of nick sites appear at least 10 times in the data set to provide adequate sampling. This reduced the total number of nick sites to 2,298,508 pairs, owing to the relatively low 36× coverage of the genome, which forced us to remove many of the nick pairs with separations greater than 300 kbp. Figure 3.2a shows that, using the 2,298,508 nick sites that were sampled at least 10 times, there is a systematic under-extension for short genomic distances between the nick sites, N_{kbp}, and low % GC values. Examining these particular nick pairs identified them as regions of the human genome containing unknown sequences (N-base regions) in the reference. We thus returned to our analysis and only included nick pairs that satisfy the above resolution criteria and do not contain any N-bases. The resulting data

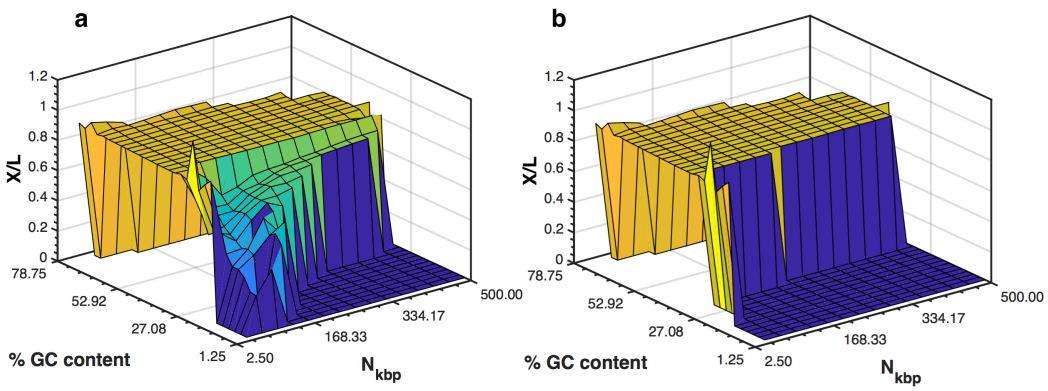

Figure 3.2: Average fractional extension as a function of % GC content and genomic distance N_{kbp} between nick sites using (a) 2,298,508 resolvable nick pairs in the human genome and (b) the 2,289,929 resolvable pairs of nick sites that do not contain any N-base regions. We chose to present the data in 3-D form, rather than as a heat map, to make it easy to visualize the effect of removing N-base regions.

set includes 2,289,929 pairs of nick sites. Figure 3.2b shows that the anomalously short extensions from the original data set are removed by eliminating the systematic error in L by only considering the known portions of the human genome.

3.4 Results

3.4.1 High-Throughput Data

Human DNA and the high-throughput afforded by genome mapping in nanochannels are essential to the robustness of our experiments. In contrast to microorganisms and viruses, whose DNA are commonly exploited for polymer physics [114], human DNA possesses a wide range of % GC content. As an extreme example, we identified pairs of nick sites with very similar separations on chromosome 6 (2,555 bp separation) and chromosome 15 (2,504 bp separation) with % GC contents of 16.4% and 74.7%, respectively. To ensure adequate sampling, we restricted our attention to % GC contents from 32.5% to 60%; each pair of nick sites in this range is sampled at least 10 times in our experiment.

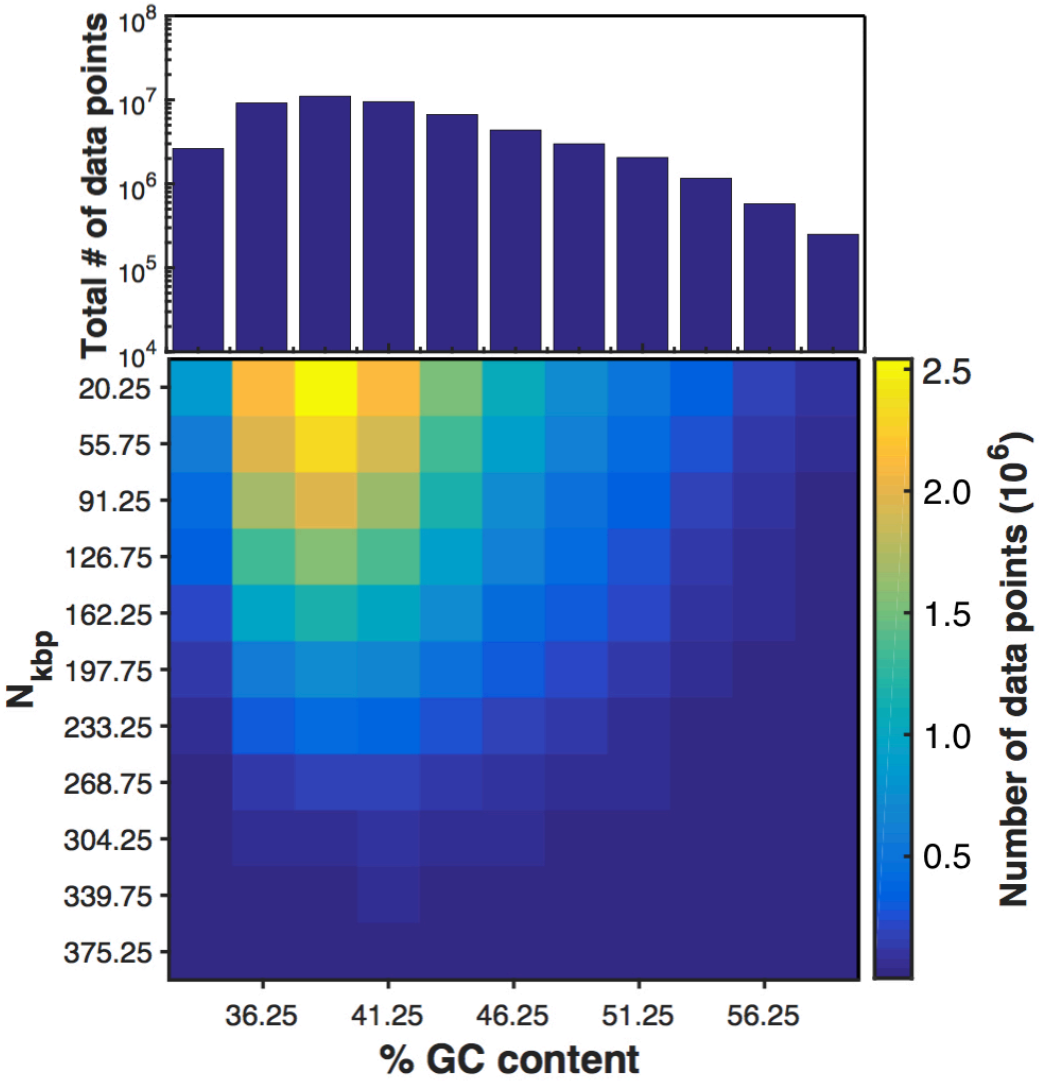

Figure 3.3: Heat map of the number of measurements of extension using a bin size of 2.5% for % GC content and 35.5 kbp for the number of kilobase pairs between nick sites, N_{kbp}. The tick labels on the left y-axis of N_{kbp} and on the bottom x-axis of % GC content indicate the midpoints of the bins. The upper histogram presents the total number of data points in each % GC content bin.

Figure 3.3 summarizes the resulting data set, which contains 50,493,547 measurements obtained from single molecules of DNA. The trend in % GC content at fixed N_{kbp} reflects the sequence of the human genome, which is AT-rich. The trend in N_{kbp} at fixed

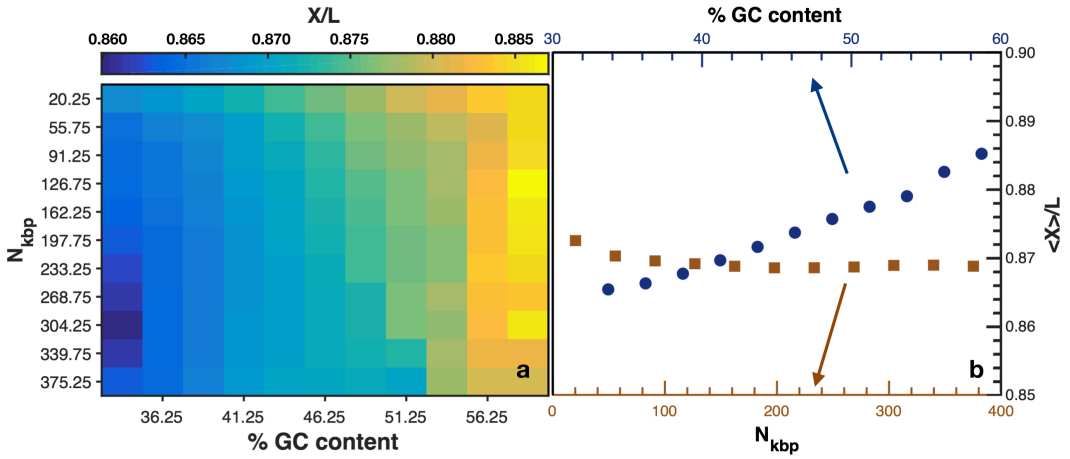

Figure 3.4: (a) Heat map of the fractional extension using a bin size of 2.5% for % GC content and 35.5 kbp for the number of kilobase pairs between nick sites, N_{kbp}. The tick labels indicate the midpoints of the bins. (b) Average fractional extension as a function of % GC content (blue circles) and N_{kbp} (brown squares). The notation $\langle X \rangle$ indicates averaging over either % GC content or N_{kbp}.

% GC content arises because each DNA molecule (e.g., Fig. 3.1d) will contribute many measurements with short distances between nick sites but only a few measurements at long distances.

Figure 3.4 shows how the fractional extension between each pair of nick sites depends on the % GC content and the genomic distance N_{kbp} between those nick sites. We report our results here in terms of the fractional extension, X/L, where X is the DNA extension measured between a pair of nick sites, assuming that the contour length L can be obtained from the 0.34 nm rise in B-DNA. While high levels of YOYO intercalation can increase L [108], the effect should be small at our low dye loading. We will address any systematic errors introduced by this assumption later.

It is reasonable to question whether the fractional extension is correlated to number of nick sites. Figure 3.5 shows the result of our analysis for the % GC content = 36.25 % bin and the N_{kbp} = 162.25 kbp bin, which we chose as a representative example. For a given pair of nick sites within that bin, we computed the number of intervening nick sites, which is the x-axis of the figure. There are 72,572 pairs of nick sites and 1,341,530 measurements in this bin, which should give a convincing statistical result as to whether

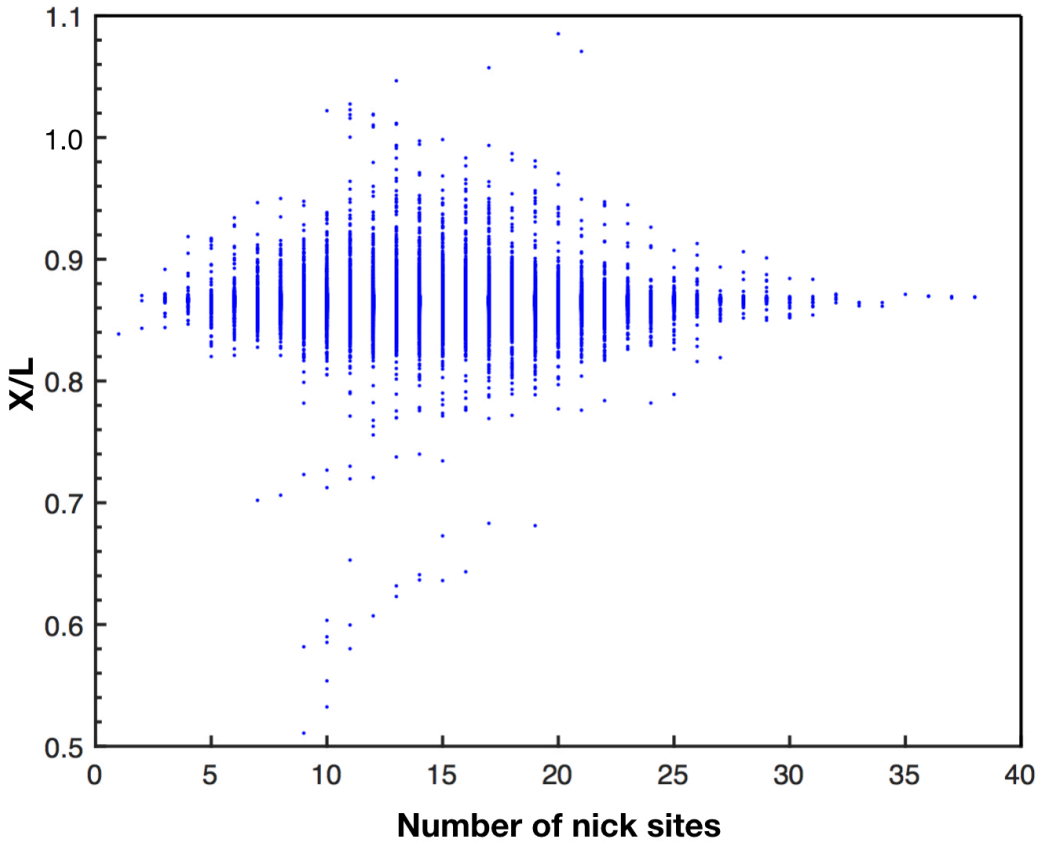

Figure 3.5: Average fractional extension as a function of the number of nick sites between a given pair of nick sites for data in the 36.25% GC content and $N_{\mathrm{kbp}} = 162.25$ kbp bin. Each data point represents the average value of X/L for pairs of nick sites in the human genome with at least 10 measurements. There are 72,572 pairs of nick sites and 1,341,530 measurements in total.

the number of nicks within a pair of nick sites is correlated with fractional extension. From visual inspection, we do not see any correlation between the fractional extension and the number of nick sites from Fig. 3.5. To be more quantitative, we did ANOVA and obtained an F-ratio $= 0.0041$. That is, we cannot reject the null hypothesis where all average fractional extension of each bin of number of nick sites are equal. There is no statistically significant correlation between the fractional extension and the number of nick sites between a given nick pair.

After excluding the potential factor, number of nick sites between a given pair of nick sites, that might affect fractional extension, we focused on the other two controlled variables, the % GC content and the genomic distance N_{kbp} between pairs of nick sites. To know which one is the significant factor to affect the fractional extension, an analysis of variance (ANOVA) and Tukey's minimum significant difference (MSD) test for the data in Fig. 3.4 are provided in the following section.

3.4.2 Statistical Analysis of Data in Fig. 3.4

Analysis of Variance (ANOVA)

Complete Data Set We first used an analysis of variance (ANOVA) to determine if there is a statistically significant increase in the fractional extension, X/L, as either the % GC content or genomic distance between nick pairs, N_{kbp}, increases. The underlying DNA sequences are 2.5 to 393 kilobase pairs long and contain % GC contents between 32.5% and 60.0%. Figure 3.6 shows how the data in Fig. 3.4a were binned using a bin size of 2.5% for % GC content and 35.5 kbp for the number of kilobase pair, N_{kbp}, between nick sites. Each data point in Fig. 3.6a and b is the average value of X/L for a given nick pair. Figure 3.6c provides a histogram for the appearance of different X/L data in a linear format, with an upper bound of $X/L = 1.4$.

Figure 3.6a and b are cutoff at $X/L = 0.5$ and $X/L = 1.4$. Bear in mind that, in our statistical analysis and calculation of the average extensions, we use all of the data obtained in the experiments (see §3.5.3 for more information and Fig. 3.11 for the full range of X/L) and weight by the number of times a nick pair appears.

For the ANOVA, the null hypothesis is

$$H_0 : \bar{X}_1 = \bar{X}_2 = ...\bar{X}_i = ... = \bar{X}_{\mathrm{k}} \tag{3.1}$$

where $\bar{X}_{\mathrm{i}}$ is the sample mean value in the i^{th} bin and k is the total number of independent bins in a comparison. The alternative hypothesis is

$$H_1 : \text{These means are not all equal.} \tag{3.2}$$

The test statistic for ANOVA is to calculate an F ratio from: (i) sum of squares

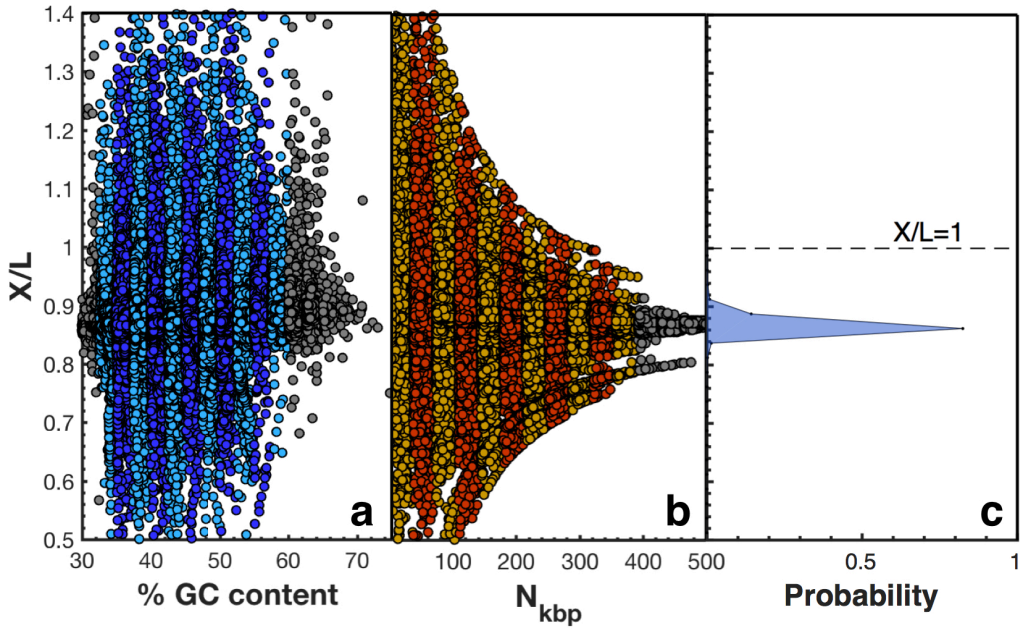

Figure 3.6: Measurements of the fractional extension X/L versus (a) % GC content and (b) genomic distance between nick pairs, N_{kbp}. Each data point represents the average value of X/L for pairs of nick sites in the human genome with at least 10 measurements. The total number of data points in these panels is 2,289,594. Light and dark blue/orange dots in Figs. 3.6a and b represent different bins using a bin size of 2.5% for % GC content and 35.5 kbp for N_{kbp}. Grey dots represent the data that were excluded from our analysis. There are a total of 11 bins in each of these panels and the plot is cutoff at $X/L = 0.5$ and $X/L = 1.4$. (c) Probability of observing an X/L value using a bin size of 0.025. The extension $X/L = 1$ is indicated by the dashed line.

within the bin (SSW),

$$\text{SSW} = \sum_{i=1}^{k} \sum_{j=1}^{N_i} (X_j - \bar{X}_i)^2 \tag{3.3}$$

where N_i is the sample size in the i^{th} bin; (ii) the degree of freedom within the bin,

$$\text{dFB} = N - k \tag{3.4}$$

where $N = 50,493,547$ is the total number of measurements and $k = 11$ is the number

		Bin number i										
Variable	Quantity	1	2	3	4	5	6	7	8	9	10	11
% GC	$N_i(\times 10^5)$	26.4	91.7	110.6	95.2	66.9	43.7	29.9	20.6	11.6	5.8	2.5
	$\bar{X}_i$	0.8654	0.8663	0.8677	0.8697	0.8716	0.8737	0.8757	0.8775	0.8790	0.8826	0.8852
$N_{\rm bp}$	$N_i(\times 10^5)$	121.3	104.6	88.3	70.8	52.2	33.6	18.5	9.0	4.1	1.8	0.7
	$\bar{X}_i$	0.8725	0.8703	0.8696	0.8691	0.8688	0.8685	0.8686	0.8687	0.8689	0.8689	0.8688

Table 3.1: Bin statistics for ANOVA. N_i is the number of entries in bin i and $\bar{X}_i$ is the average value in bin i.

of bins; (iii) the sum of squares between the bins (SSB),

$$\mathrm{SSB} = \sum_{i=1}^{k} N_i(\bar{X}_i - \bar{X})^2 \tag{3.5}$$

where $\bar{X} = 0.8701$ is the overall mean; and (iv) the degree of freedom between the bins,

$$\mathrm{dfB} = k - 1 \tag{3.6}$$

For notational simplicity, we use the variable X here to denote the fractional extension (X/L). The F-ratio is then given by

$$F = \frac{\mathrm{SSB/dfB}}{\mathrm{SSW/dfW}} \tag{3.7}$$

The statistics for N_i and $\bar{X}_i$ are in Table 3.1. The bin-averaged fractional extensions reported in Table 3.1 appear as the values of $\langle X\rangle/L$ in Fig. 3b.

We obtained $F = 29.60$ for % GC content and $F = 0.98$ for $N_{\rm kbp}$. These ratios indicate that the means for X/L binned by % GC content are more significantly different than the means binned by $N_{\rm kbp}$. The decision rule for an ANOVA F-test is to compare the test statistic result with an appropriate critical F value determined by dfW and dfB from a table of probabilities for the F distribution. To reject the null hypothesis at a 95% confidence level (which we also use later for Tukey's minimum significant difference), the test statistic needs to be larger than the critical value, $F_c = 1.83$. That is, we only reject H_0 for the comparison that was binned by % GC content. The value of $F = 29.60$ for % GC content even rejects the null hypothesis at a 99.99999999999999% confidence level ($F_c = 9.98$) calculated by a Matlab F inverse cumulative distribution

function (`finv`).

Re-sampled Data There are correlations in our data that are not included in the preceding ANOVA analysis because each molecule contributes many barcode pairs to the data set. To see if the correlations affect our conclusion, we implemented a resampling method to our data set to minimize the possible systematic error. Starting with the 2,289,929 pairs of nick sites in the data set, we randomly picked a set of nick sites pairs in which any pair of nick sites is at least 500 kbp away from any other pair of nicking sites in the resampled data. This restriction is trivially satisfied by nick pairs on different chromosomes, so the restriction was only imposed between pairs of nick sites on the same chromosome. By doing so, we could dramatically reduce the probability that a pair of nick sites came from the same molecule; only molecules that can span this large gap *and* have both of the randomly selected nick pairs on either side of the gap could be correlated. This is highly unlikely. We repeated this analysis 100 times, each time randomly picking a new set of nick pairs, and there were around 3,000 pairs of nick sites in each resampled data set. ANOVA was applied to calculate the F-ratio for each of these 100 data sets after binning them by N_{kbp} and % GC content, respectively.

The critical value for these 100 resampled data sets is still $F_c = 1.83$. Figure 3.7 shows the F-ratio computed for each of the 100 resampled data sets. For clarity, panel (a) shows the F-ratios, while panels (b) and (c) show the ratios of the F-ratios for each re-sampling of the data. While the F-ratio values binned by N_{kbp} now are higher than the critical value due to the smaller degree of freedom within the bin, the F-ratio values which were binned by % GC content are much higher than the values binned by N_{kbp}, which indicates that there is larger difference between $\langle X \rangle / L$ binned by % GC content than binned by N_{kbp}. The resampled ANOVA leads to the same conclusion as our first attempt of analysis, namely that we should bin our data by % GC content instead of N_{kbp}.

Tukey's Minimum Significant Difference (MSD)

We also used Tukey's minimum significant difference (MSD) to find means that are significantly different from each other in both of the comparisons. The MSD is defined

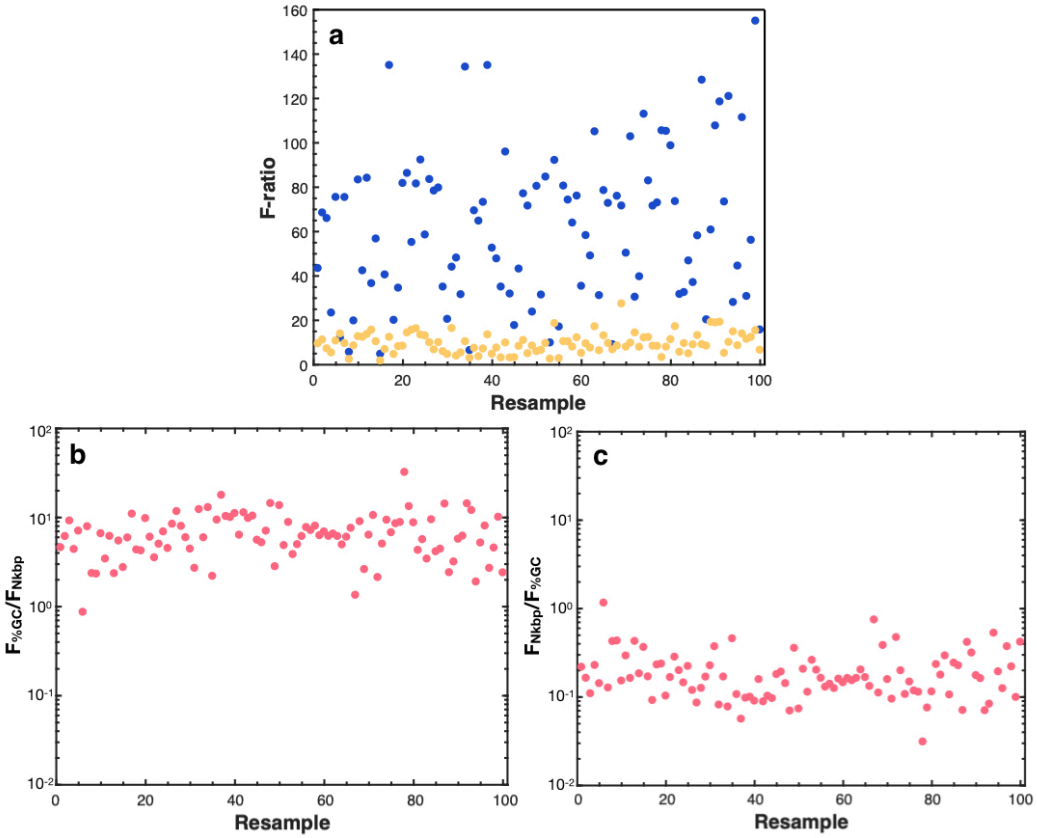

Figure 3.7: (a) F-ratio of 100 resampled data sets obtained by ANOVA. Blue circles are the F-ratios for data binned by % GC content, and yellow circles are F-ratios for data binned by N_{kbp}. There are around 3,000 pairs of nick sites in each resampled data set, and each pair of nick sites is at least 500 kbp away from others if they are on the same chromosome. (b) The ratio of F-ratios binned by % GC content to N_{kbp}. (c) The ratio of F-ratios binned by N_{kbp} to % GC content.

as

$$\mathrm{MSD} = Q_{\alpha(k,N-k)}\sqrt{\frac{(\mathrm{SSW}/\mathrm{dfW})}{N^*}} \tag{3.8}$$

where Q is a critical value determined by a significance level, α, found in a Studentized Range q Table, and N^* is a modified sample size,

$$N^* = \frac{k}{\sum_{i=1}^{k} N_i^{-1}} \tag{3.9}$$

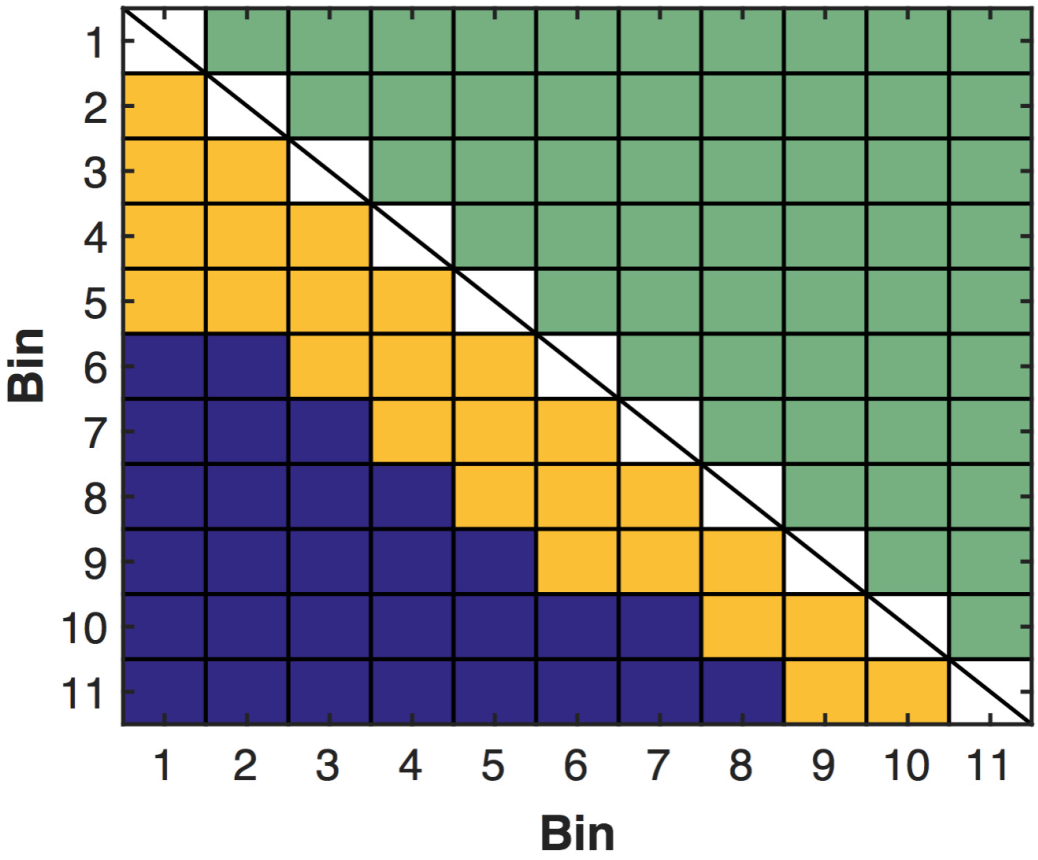

Figure 3.8: Results of Tukey's MSD test at an $\alpha = 0.05$ significance level. In the figure, the area below the diagonal is the comparison of % GC content, and the upper part is the comparison of N_{kbp}. Each box represents the result of an MSD test which compares the means for X/L of the corresponding column bin i and row bin j. Boxes labeled in green indicate that the means are not significantly different when binned by N_{kbp}; boxes in yellow indicate that the means are not significantly different when binned by % GC content, and boxes in blue indicate that the means are significantly different when binned by % GC content. The diagonal of the plot, corresponding to MSD between the same bin, is not meaningful.

At a significance level $\alpha = 0.05$ (i.e., 95% confidence), corresponding to $Q = 4.43$, we obtain MSD = 0.0064 for % GC content and MSD = 0.0224 for N_{kbp}. If the absolute value of the difference of either two means of bins, $|\bar{X}_i - \bar{X}_j|$, in a comparison is larger than the MSD, then the means of these two bins are significantly different from each

other at the corresponding significance level.

Figure 3.8 summarizes the result of Tukey's MSD test for a significance level $\alpha = 0.05$. We can see all pairs of means in the N_{kbp} comparison fail the MSD test, which is consistent with the result of ANOVA. That is, the results from ANOVA and MSD both indicate that means for X/L are more significantly different when binned by % GC content.

Based on the statistical analysis, we conclude that the increase in the average factional extension, $\langle X\rangle/L$ as % GC content increases (blue circles in Fig. 3.4b) is statistically significant. In contrast, $\langle X\rangle/L$ when binned by N_{kbp} is not statistically different (brown squares in Fig. 3.4b).

We thus proceed by only binning the data with respect to % GC content. Figure 3.4b shows that the change of the average fractional extension, $\langle X\rangle/L$ versus % GC content is small, around 2%. However, this small change is crucial to our genomic strategy. Genome mapping is required to obtain the measurements of L from the DNA sequence. The mapping method is robust to such small changes in extension since it is a *de novo* method that relies on pattern recognition [115]. Drawing a statistically meaningful conclusion, though, requires precise measurements of $\langle X\rangle/L$. Figure 3.3 indicates that each of these % GC content bins contains between 10^5 to 10^7 measurements. As a result, the standard error of the average extension, $\langle X\rangle$, within a given % GC content bin is very small.

3.5 Discussion

3.5.1 Statistical Terpolymer Model

Simulations of channel-confined wormlike chains [51,100] indicate that, for the fractional extensions in Fig. 3.4b, the chain lies within the Odijk regime [93]. The corresponding fractional extension is predicted to be [93, 95]

$$\langle X\rangle/L = 1 - 0.18274(D_{\mathrm{eff}}/l_{\mathrm{p}})^{2/3} \tag{3.10}$$

where D_{eff} is the effective channel size available to the chain. For very small channels, such as those used here, the exact value of D_{eff} is not obvious due to the electrostatic

interactions between DNA and the channel walls [44, 116]. However, we would expect those interactions to be independent of sequence. To proceed, we adopt the standard approximation [100] of $D_{\text{eff}} = D - w$, where $w = 7.6$ nm is the Stigter effective width [75] for our 48 mM buffer. As was the case with L, we will address any systematic errors from this assumption shortly. Inverting Eq. (3.10) yields the persistence length, and the results as a function of % GC content are shown as blue solid circles in Fig. 3.9.

The sequence-dependence of the DNA persistence length can be explained by modeling the DNA as a statistical terpolymer, illustrated in the inset of Fig. 3.9 and described in more detail in Appendix A. The particular sequence of the DNA is replaced by an effective sequence where a G-C bond is replaced by S (strong hydrogen bonding) and an A-T bond is replaced by W (weak hydrogen bonding). The bending energy depends not on each base itself but on the sequence of dinucleotide pairs [68]: E_{SS}, E_{SW}, and E_{WW}. Previously, Hogan *et al.* measured these bending energies by triplet state anisotropy decay [68]. We constrain the present model by the ratio of the bending energies obtained in these experiments: $E_{\text{SW}}/E_{\text{SS}} = 1.4/2.9$ and $E_{\text{WW}}/E_{\text{SS}} = 0.82/2.9$ [68]. The persistence length at large length scales emerges from the local bending energies. As such, the relevant bending energy is the weighted average of the dinucleotide pairs in the sequence,

$$E = \sum_{i,j} p_{ij} E_{ij} \tag{3.11}$$

where $(i, j) \in (\text{S}, \text{W})$. Denoting the % GC content (i.e., the probability of locating a G or C base) by γ, the probabilities p_{ij} of observing particular dinucleotide pairs in a statistical terpolymer are $p_{\text{WW}} = (1-\gamma)^2$, $p_{\text{SW}} = p_{\text{WS}} = \gamma(1-\gamma)$ and $p_{\text{SS}} = \gamma^2$, leading to the bending energy $E = E_{\text{WW}}(1-\gamma)^2 + E_{\text{SW}}\gamma(1-\gamma) + E_{\text{SS}}\gamma^2$. Assuming that the surface moment of inertia, I_{s}, is independent of the sequence, the intrinsic persistence length is given by $l_{\text{p},0} = EI_{\text{s}}/k_{\text{B}}T$ [68]. Polyelectrolyte theory [52, 53, 67] further requires that the persistence length include an electrostatic contribution $l_{\text{p,el}}$ due to screening of backbone charges by the counterions in solution. We assume that all sequences are affected by electrostatics in the same manner, since they arise from the acidic backbone. By fitting to experimental data for λ-DNA, Dobrynin [67] obtained

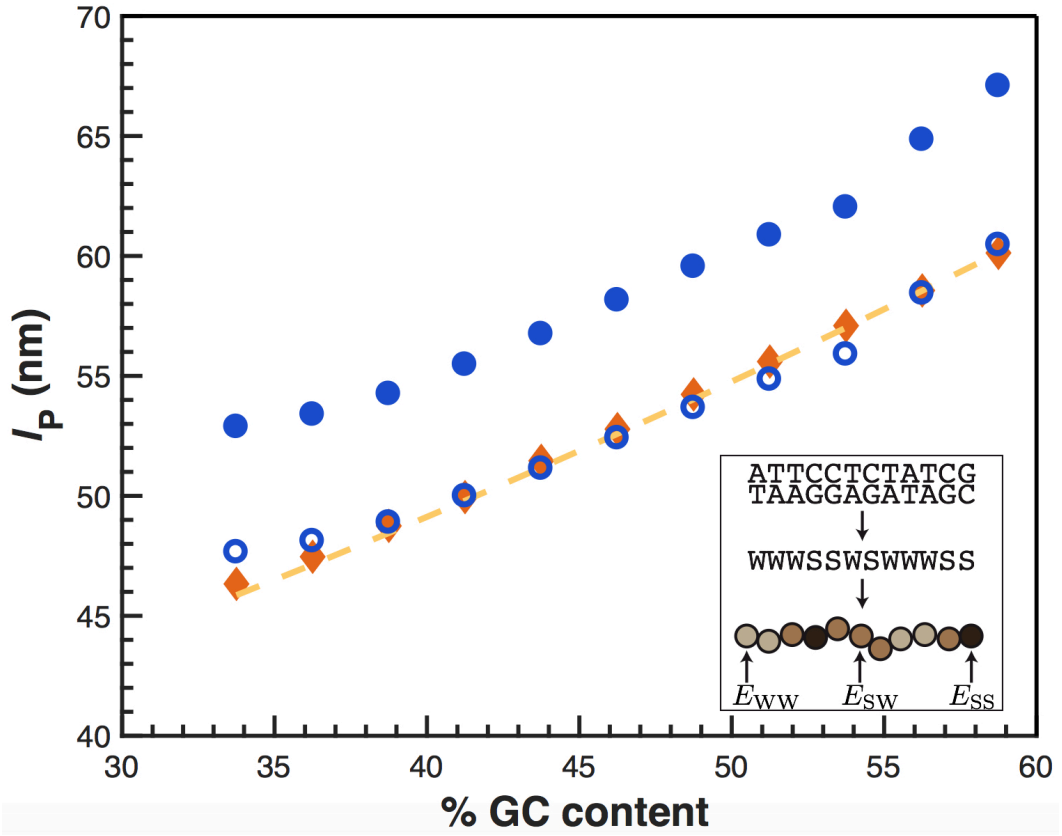

Figure 3.9: Persistence length as a function of % GC content. Blue circles are experimental data using $D_{\text{eff}} = D - w = 33.4$ nm (solid circles) and $D_{\text{eff}} = 30.1$ nm (open circles). Dashed line is the statistical terpolymer model prediction in Eq. (3.13) and the orange diamonds are the model predictions using the average dinucleotide composition in each % GC content bin. Inset: Statistical terpolymer model. The DNA sequence is converted first into a sequence of strong (G-C) and weak (A-T) hydrogen bonds. The persistence length is computed from the resulting sequence of dinucleotide pairs (WW, SW, or SS) based on their respective bending energies E_{ij}, where $i, j =$ (S,W).

the empirical formula

$$l_{\mathrm{p}}\ [\mathrm{nm}] = l_{\mathrm{p},0} + l_{\mathrm{p,el}} = 46.1 + \frac{1.9195}{\sqrt{I\ [\mathrm{M}]}} \tag{3.12}$$

where I is the ionic strength. Using $\gamma = 0.4986$ for λ-DNA yields $E_{\mathrm{SS}} I_{\mathrm{s}}/k_{\mathrm{B}}T = 82.2$ nm (see Appendix A). As a result, the statistical terpolymer model predicts (see Appendix A)

$$l_{\mathrm{p}}\ [\mathrm{nm}] = (23 + 33\gamma + 26\gamma^2) + \frac{1.9195}{\sqrt{I\ [\mathrm{M}]}} \tag{3.13}$$

This is the key result of our analysis, and extends Dobrynin's result for the GC-even genome of λ-phage DNA to the range of sequences commonly found in human DNA.

Figure 3.9 shows that Eq. (3.13) (dashed line) captures the trend in persistence length as a function of % GC content. As noted previously, there are systematic errors due to the intercalation of YOYO dye (which affects L) and the DNA-wall electrostatic interactions (which affect D_{eff}). It is also possible that there is an additional source of systematic error from the effect of intercalation on the persistence length, but there is a growing body of systematic experimental work [60,87] indicating that intercalation does not affect the persistence length. These systematic errors should affect all sequences in the same manner, so they would shift the prediction of the model up or down, but would not change the curvature. Indeed, Fig. 3.9 shows that we can bring the model into agreement with the experiments by assuming $D_{\mathrm{eff}} = 30.1$ nm (open circles in Fig. 3.9), which is certainly within reason based on the uncertainty in the DNA-wall interactions [44,116] and the accuracy of the SEM characterization of such a large array of channels.

To check the accuracy of assuming a random sequence, we also computed the dinucleotide composition between pairs of nick sites from the DNA sequences that lie within a given % GC content bin, and then recomputed the predictions of the model by replacing the probabilities in Eq. (3.11) with those data. Figure 3.9 shows that accounting for the exact DNA sequence (orange diamonds in Fig. 3.9), rather than assuming a random sequence with a particular averaged % GC content (dashed line in Fig. 3.9), hardly affects the result.

On the basis of the result we demonstrated, we were interested how our conclusion

would be affected if more restrictions were applied to our data. We also examined whether the accuracy of the model could be improved with 10-dinucleotide model of Geggier *et al.* [2]. In the following sections, we discuss several cases where our data was treated with additional analyses first and compared to the model afterwards.

3.5.2 Results with the Data Binned by Variable Width Bins

Table 3.1 shows the results of bin-averaged fractional extensions binned by equal width bins in % GC content and N_{kbp}. However, the results in bins of high % GC content and high N_{kbp} have significantly less data than other bins. To check whether this difference in bin occupancy number affects our result, we binned our data with variable width bins and repeated the analysis. However, given the amount of data we acquired and the way it was compiled, it is prohibitively costly to re-sort all of the data to have exactly the same number of measurements in each bin. As a more computationally feasible alternative, we first re-binned our data with a very small bin size of % GC content, 0.1 %. We then clustered these bins and so that there are approximately 200,000 to 600,000 measurements in each % GC content bin.

Figure 3.10 shows that the result for the persistence length with variable width bins (light blue open circles) is essentially the same with the data binned by uniform width bins (blue solid circles). We conclude from this figure that the different bin occupancy numbers did not affect our analysis.

3.5.3 Outliers in the Data Set

Semilogarithmic Plot of Fig. 3.6c

Figure 3.6c provides a histogram for the appearance of different X/L data in a linear format, with an upper bound of $X/L = 1.4$. Figure 3.11 provides the same result in a semilogarithmic format with an upper bound of $X/L = 5$. Figure 3.11 does a better job of presenting the outliers than Fig. 3.6c, as they are so rare that they are smaller than the line thickness when plotted in a linear format in Fig. 3.6c. However, Fig. 3.11 requires some care in interpreting the importance of the outliers since their frequency is several orders of magnitude lower than the center of the distribution but the semilogarithmic of the plot makes them seem more frequent at first glance.

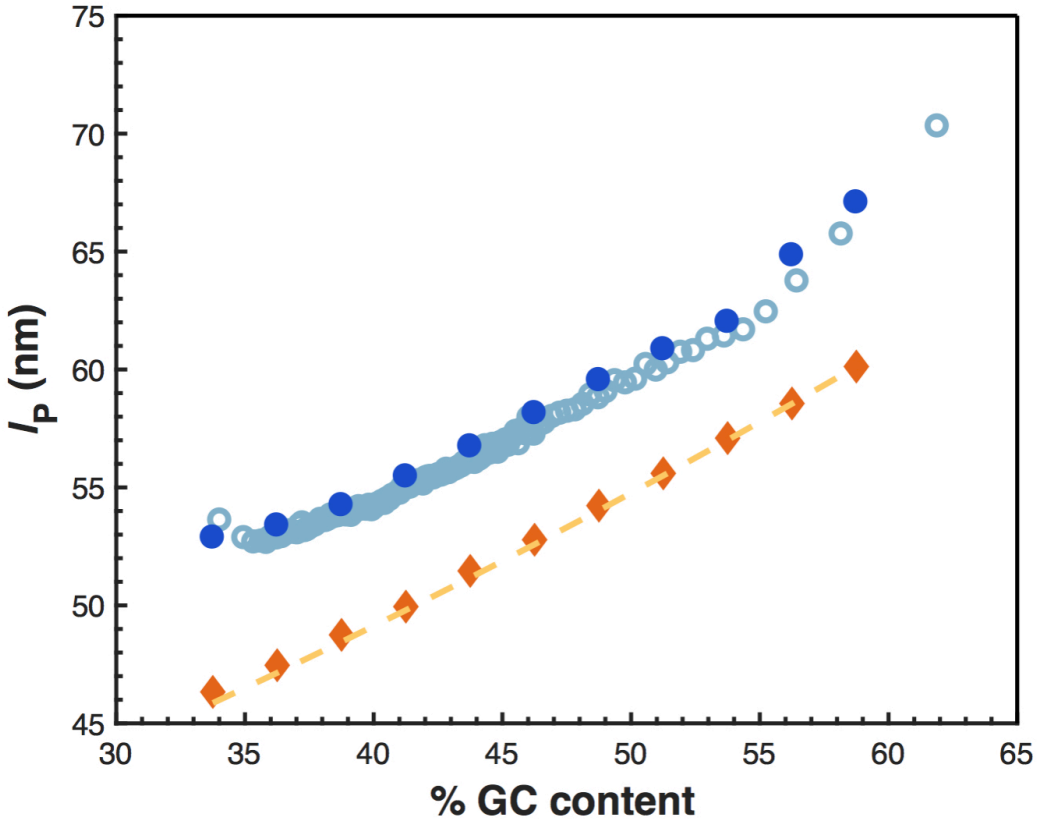

Figure 3.10: Persistence length as a function of % GC content. Blue circles are experimental data binned by the consistent width bins (solid circles) and binned by the variable width bins (light blue open circles). Dashed line is the statistical terpolymer model prediction in Eq. (3.13) and the orange diamonds are the model predictions using the average dinucleotide composition in each % GC content bin with consistent width bins.

The physically impossible values of X/L arise from the systematic errors in L due to differences between the genome of the cell line and the human reference genome. While these errors occur hundreds of times, they still only represent a very small portion of the data set. Note that a single DNA molecule that covers a structural variation with respect to the reference human genome can contribute many times to this histogram since each pair of nick labels on that molecule constitutes a single measurement.

The number of such anomalous measurements decreases as N_{kbp} increases due to

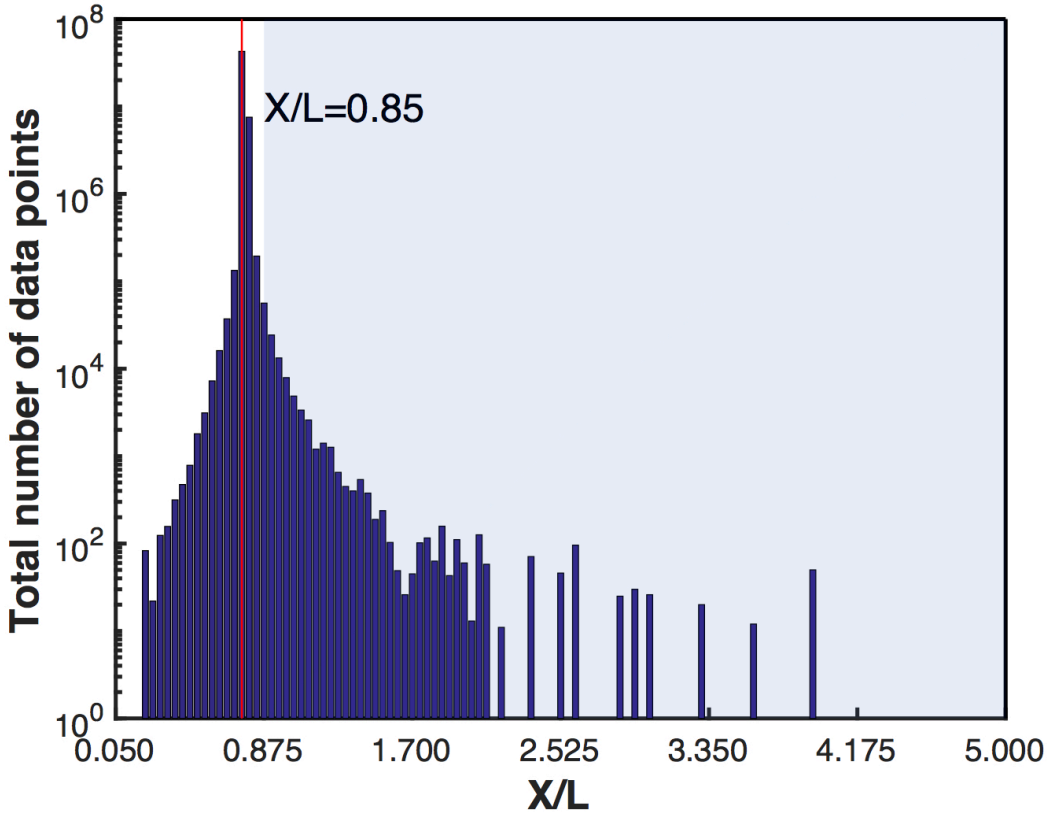

Figure 3.11: Histogram of X/L in semilog version using a bin size of 0.05 for X/L. Most of the data are strongly concentrated around the mean extension $X/L = 0.87$. The light blue shading represents the area where $X/L > 1$. The maximum in the distribution is at $X/L = 0.85$.

dilution; a small structural variation can produce a large error in L for small N_{kbp} but its effect is minimal for large N_{kbp}. However, it is more likely that the effect can be explained by the relationship between the standard deviation of fractional extension and contour length. Explicitly, the extension X scales like L while its variance scales like $L^{1/2}$. It then follows that the standard deviation of the fractional extension has the scaling

$$\langle (X/L - \bar{X}/L)^2 \rangle^{1/2} \sim L^{-1/2} \tag{3.14}$$

Figure 3.12 shows the heat map of the standard deviation of fractional extension binned

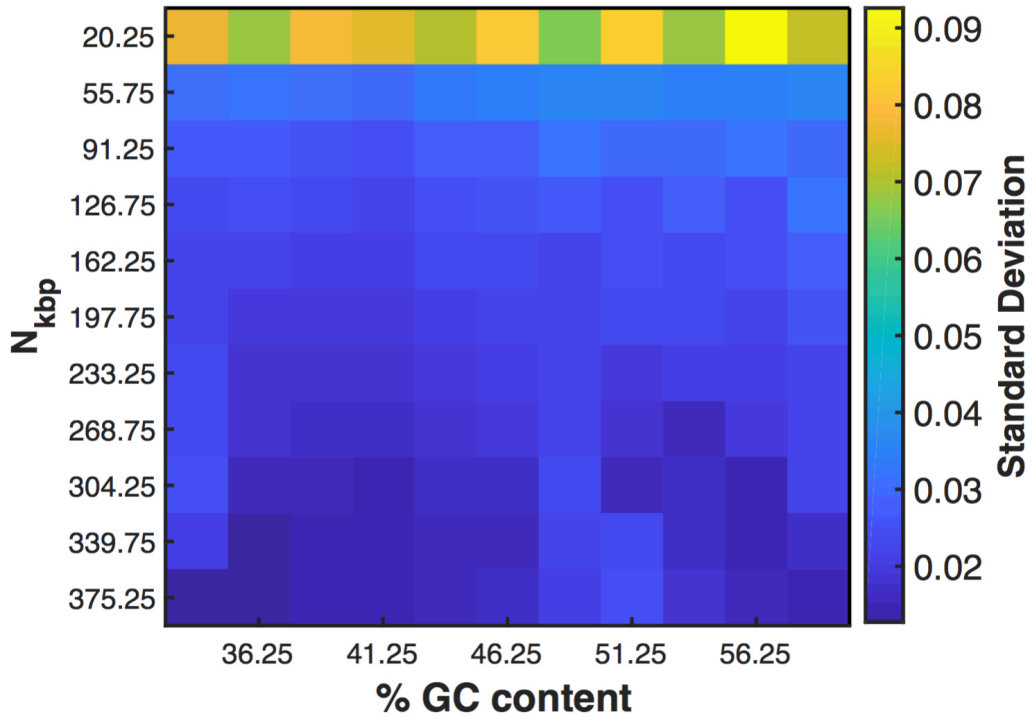

Figure 3.12: Heat map of the standard deviation of fractional extension using a bin size of 2.5% for % GC content and 35.5 kbp for the number of kilobase pairs between nick sites, N_{kbp}. The tick labels indicate the midpoints of the bins.

by % GC content and N_{kbp}. The standard deviation of fractional extension for the smallest N_{kbp} bin is much higher than the standard deviations for other values of N_{kbp}, a result which is expected from Eq. (3.14). The higher standard deviation of fractional extension at short N_{kbp} can be due to the structural variation, as we discussed previously, or it can also be due to the effect of different sequence orders, such as AAGG, or AGAG, which will have larger influence on shorter N_{kbp} than longer N_{kbp}. Fortunately, Fig. 3.6a shows that the anomalously large measurements of X/L are distributed throughout the % GC content range and thus cannot be the source of the statistically significant increase in X/L as a function of % GC content.

We thus chose to work solely in terms of % GC content and include in our analysis all values of N_{kbp}. In principle, additional bioinformatic analysis would allow us to

determine the location of the structural variations within the genome of this cell line, and thus allow us to remove those nick pairs that contain a systematic error in L due to the structural variation. Moreover, removing the structural variations would also further flatten the curve for $\langle X \rangle / L$ as a function of N_{kbp} in Fig. 3.4b, since structural variations play a stronger role at small values of N_{kbp}. However, given that almost all of the data lie within physically reasonable values of X/L and the weak trend in $\langle X \rangle / L$ as a function of N_{kbp} has low statistical significance, there is little to be gained by removing the thousands of points (amongst fifty million) that are outliers.

Result after Imposing Additional Quality Cuts on the Data

Although these outliers are just a small portion of the data set, it is worthwhile to determine if the final result would be improved by imposing additional quality cuts on the data. To test this question, we extracted data with the range of fractional extension, [0.825, 0.950], which was determined by Fig. 3.6c. Figure 3.13 shows that our statistical terpolymer model prediction in Eq. (3.13) (orange dashed line) is even better after imposing additional quality cuts (black open circle) at high % GC content. However, the choice of the values of X/L to cut the data is somewhat arbitrary. As a result, we use all of the data for the figures in the dissertation research and our statistical analysis.

3.5.4 Comparison of Experimental Results with the 10-Dinucleotide Model of Geggier *et al.* [2]

In Fig. 3.9, we compare our experimental results with a statistical terpolymer model that only accounts for the bending energy between three types of dinucleotide pairs: (i) strong/strong, (ii) strong/weak or weak/strong, and (iii) weak/weak, where the "strong" base pairs are G/C and the "weak" base pairs are A/T.

Geggier *et al.* [2] developed a more sophisticated model that takes into account the ten distinct types of dinucleotide pairs listed in Fig. 3.14. Based on symmetry properties of the DNA bases [117], Geggier *et al.* [2] report persistence lengths for each dinucleotide pair in Table S2 of their paper.

In order to compare our experimental results with that predicted by the model of Geggier *et al.* [2], we calculated the average dinucleotide compositions. For each % GC

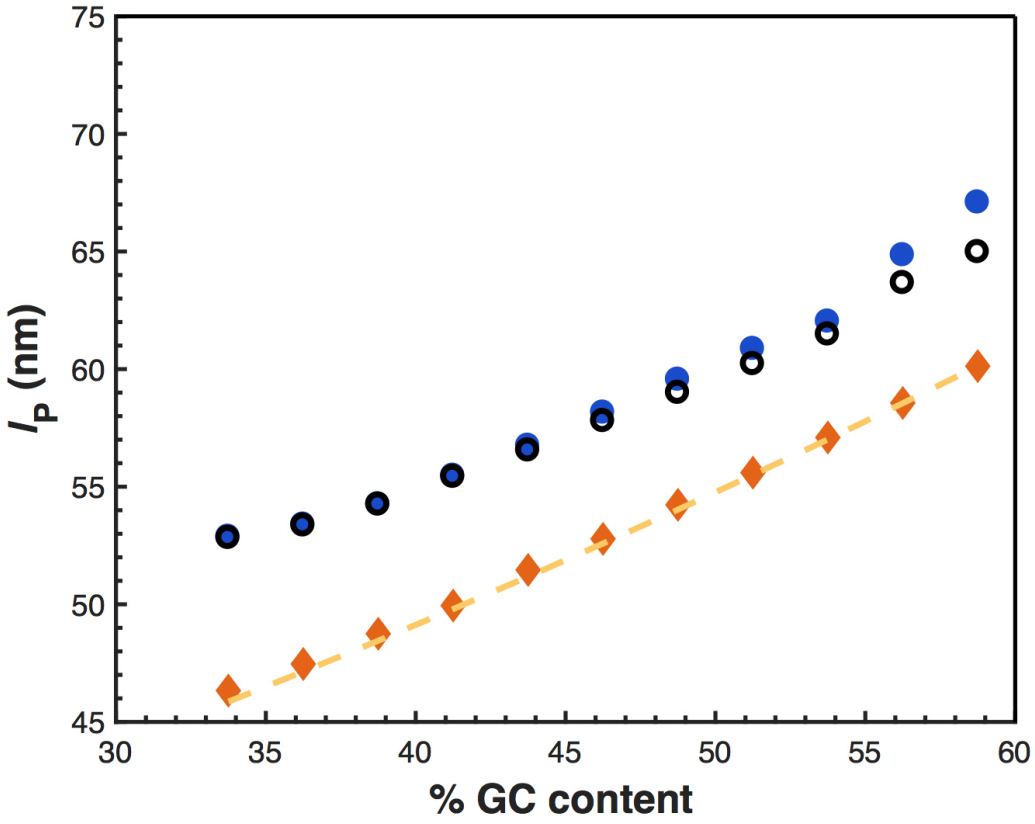

Figure 3.13: Persistence length as a function of % GC content. Blue solid circles are experimental data without additional quality cuts, and black open circles are experimental data with additional quality cuts within the range of fractional extension, [0.825, 0.950]. Dashed line is the statistical terpolymer model prediction in Eq. (3.13) and the orange diamonds are the model predictions using the average dinucleotide composition in each % GC content bin.

content bin, we first calculated average dinucleotide composition for each N_{kbp} sub-bin therein. By way of example, Fig. 3.14a shows how the dinucleotide composition changes with different N_{kbp} within the % GC content = 46.25% bin. The average fraction for each dinucleotide step is essentially unchanged by the number of kilobase pairs N_{kbp} between the nick sites. As a result, we can estimate the dinucleotide composition for a given % GC content bin by the result of any bin of N_{kbp} therein.

We also performed the same analysis by examining bins in N_{kbp} and calculating the

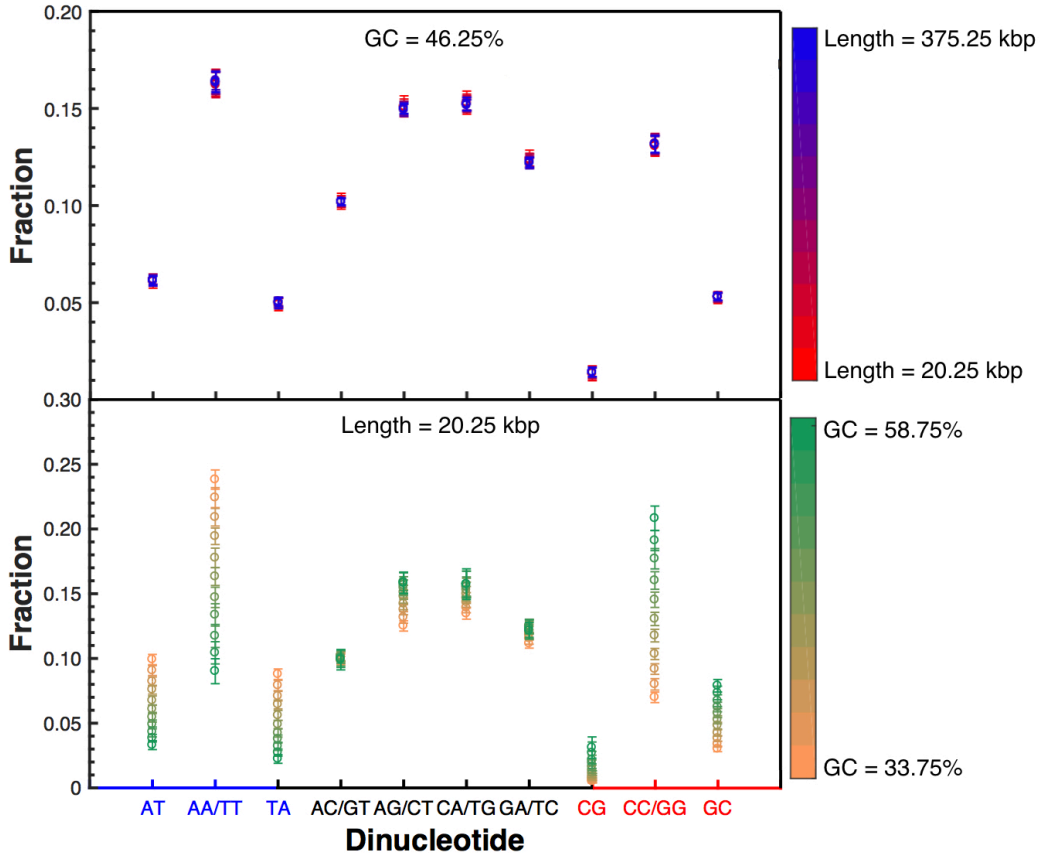

Figure 3.14: Fractions of different dinucleotide steps for (a) different N_{kbp} at % GC content = 46.25% and (b) different % GC content at N_{kbp} = 20.25 kbp. The model of Geggier *et al.* [2] uses all ten dinucleotide pairs to compute the bending energy of DNA. To provide correspondence with the statistical terpolymer model, the weak/weak dinucleotide pairs are blue, the weak/strong and strong/weak pairs are black, and the strong/strong pairs are red.

average dinucleotide composition for each % GC content sub-bin therein. Figure 3.14b shows the result of average dinucleotide composition with various % GC content at N_{kbp} = 20.25 kbp. Clearly, the composition varies and does so in the expected manner. For example, the weak/weak dinucleotide fractions increase as % GC content decreases.

We estimated the average persistence length of different % GC content by a weighted

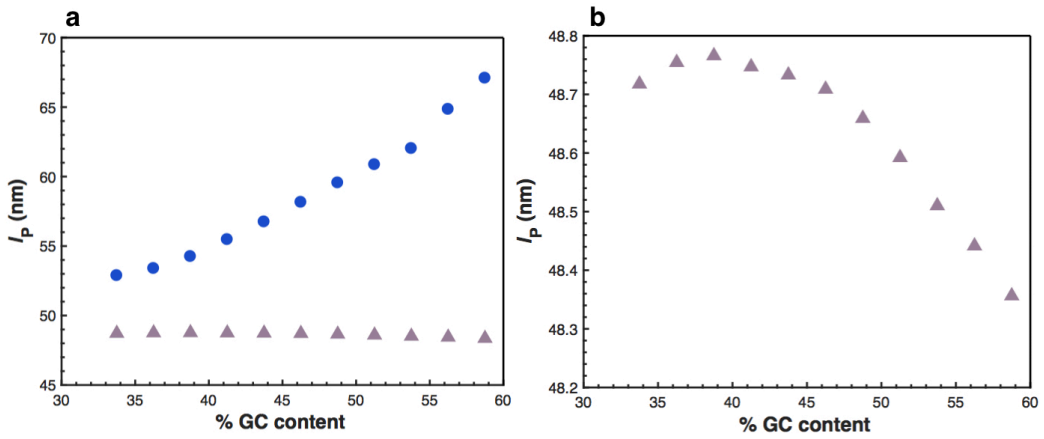

Figure 3.15: Persistence length as a function of % GC content. (a) Blue circles are experimental data using $D_{\text{eff}} = 33.4$ nm, and grey triangles are the prediction of Eq. (3.15). (b) Same as (a) with a smaller range for the y-axis.

average,

$$l_{\text{p}} = \sum_{i=1}^{10} f_i l_{\text{p},i} \tag{3.15}$$

where f_i is the fraction of the ten dinucleotide pairs in Fig. 3.14b appearing in that % GC bin and $l_{\text{p},i}$ is the persistence length of that dinucleotide pair in Table S2 of Ref. [2]. The form of Eq. (3.15) is the same as that used in our statistical terpolymer model, but now we use the persistence length data from Geggier *et al.* [2]. The result of this calculation is shown in Fig. 3.15. There is a slightly negative correlation between our experimental data and the estimation from Eq. (3.15), and the dynamic range in l_{p} is much smaller than what is observed in the experiments.

The disagreement between the nanochannel stretching experiments and Eq. (3.15) is expected from the complexity of the human genome. As noted by Geggier *et al.* [2] in their study, the persistence lengths obtained from their cyclization experiments are not expected to be a good model for sequences which contain poly(A) tracts, $\text{poly(A)}_n{\cdot}\text{poly(T)}_n$, and/or GGGCCC motifs that possess substantial intrinsic curvature [112,118–122]. Those sequence elements were all excluded from the fragments used to obtain the data reported in Ref. [2], but they are scattered in the human genome and should affect the overall persistence length in our study. It would be interesting

in future work to repeat the analysis in the present contribution using only nick sites separated by sequences appearing in the data set of Geggier *et al.* [2] to probe further the utility of their model.

One untested assumption in our model is the exclusive incorporation of sequence effects into the intrinsic persistence length. It is relatively straightforward, albeit tedious, to test this assumption by repeating the present experiments at different ionic strengths [123, 124]. We are optimistic that such experiments will validate Eq. (3.13), as previous experiments on confined DNA [124] provide convincing evidence that the dependence on ionic strength is correct and electrostatic interactions should govern long-range interactions.

3.6 Concluding Remarks

Using a high-throughput genome mapping approach, we have obtained circa 50 million measurements of the extension of internal human DNA segments in a 41 nm $\times$ 41 nm nanochannel. The underlying DNA sequences, obtained by mapping to the reference human genome, are 2.5 to 393 kbp long and contain % GC contents between 32.5% and 60%. Using Odijk's theory for a channel-confined wormlike chain, these data reveal that the DNA persistence length increases by almost 20% as the % GC content increases. The increased persistence length is rationalized by a model, containing no adjustable parameters, that treats the DNA as a statistical terpolymer with a sequence-dependent intrinsic persistence length and a sequence-independent electrostatic persistence length. We have demonstrated that the persistence length of long DNA has a remarkable dependence on the underlying sequence. We are optimistic that the model proposed in Eq. (3.13) will prove useful for quantitative analysis of DNA-based experiments.

Chapter 4

Extension distribution for DNA confined in a nanochannel near the Odijk regime

This chapter is based on the publication
H.-M. Chuang, J. G. Reifenberger, A. B. Bhandari, and K. D. Dorfman, "Extension distribution for DNA confined in a nanochannel near the Odijk regime" *J. Chem. Phys.* **151**, 114903 (2019). [125]

4.1 Introduction

When a DNA molecule is confined in a channel or pore, the excluded volume [48] and the stiffness of the polymer chain [93] cause the molecule to extend along the channel axis. It is particularly challenging to describe the thermodynamics of this process when the channel size, D, is commensurate with the persistence length, l_p, of the polymer. For $D \ll l_\mathrm{p}$, known as the Odijk regime [93], the chain is extended close to its contour length, L, with small fluctuations in alignment with respect to the channel axis [49, 94, 95]. Conversely, for $D \gg l_\mathrm{p}$, the polymer can easily form hairpin bends; the chain statistics here follow blob theory, described first by the marginal solution behavior embodied in the extended de Gennes regime [50,100,126,127] and eventually cross-over to the de Gennes

regime for $D \gg l_p^2/w$ [48]. While Odijk [49] proposed an additional scaling regime between the classic Odijk behavior [93] and the blob regimes [48], observing the scaling behavior in this so-called backfolded Odijk regime requires that the polymer exhibit a very large monomer anisotropy [51], the latter measured by the ratio of the persistence length to the effective width, w, of the polymer. For DNA, the monomer anisotropy l_p/w saturates at $l_p/w \approx 10$ at high ionic strength and decreases with decreasing ionic strength [64]. As a result, DNA will not exhibit a backfolded Odijk regime [51]. Rather, the extension of DNA confined in channels near the persistence length typically appears as a smooth transition from the de Gennes scaling to the Odijk scaling, with an apparent extension scaling that is inverse in the channel size [96–101].

Recently, two complementary theories have emerged that subsume the physics of the extended de Gennes regime and the backfolded Odijk regime into a single regime characterized by weak excluded volume interactions [4, 102]. The key result of both theories [4, 102] is the existence of a new scaling parameter, α, that represents the typical number of overlaps per hairpin bend in the chain, with the fractional extension of the chain scaling like $\alpha^{1/3}$. Importantly, the existence of the latter scaling law is not predicated on a large value of the monomer anisotropy. As a result, these theories [4,102] should permit a description of DNA extension when $D \approx l_p$, a technologically important case [29] for which the inequalities required in previous scaling theories [49] cannot be satisfied.

DNA is a polyelectrolyte, and nanochannel confinement of DNA takes place in a system with charged walls. The varied electrostatic interactions in the physical system are captured within the prevailing neutral wormlike chain model by computing an effective width w [75], which accounts for the electrostatic contribution to the segmental excluded volume, a persistence length l_p that includes contributions from both the intrinsic stiffness and the electrostatic repulsion along the DNA backbone [67], and an effective channel size D_{eff}, which accounts for the region of the channel that is inaccessible to the DNA due to DNA-wall electrostatic interactions [44, 83, 100]. Once this mapping is complete, the DNA extension within the nanochannel can be described using one of the myriad theories [4, 48, 49, 93, 102] developed for the confinement of a neutral polymer between hard walls.

The present contribution addresses the applicability of this neutral wormlike chain

model to describe the distribution of DNA extensions for channel sizes somewhat smaller than the persistence length, and thus being proximate to the Odijk regime but not yet satisfying the strong inequality $D \ll l_{\mathrm{p}}$. We take advantage of one of the new theories, the weakly-correlated telegraph model [4], to facilitate this analysis. Recently, Ödman *et al.* have derived an asymptotic solution to the telegraph model for such small channels, and proposed a method to augment that result to account for the effect of alignment fluctuations [128]. This theory is reviewed in detail in §4.2. The resulting probability distributions for the extension X of the chain are in remarkably good agreement with direct simulations of a confined, neutral wormlike chain [3] when the variance due to alignment fluctuations corresponds to that in the Odijk regime [95]. The agreement between simulation and theory persists down to values of $\alpha \approx 0.3$, well below the asymptotic requirement $\alpha \gg 1$ in the theory [3]. The predictions of the telegraph model [128] have also been compared to four of the experimental data points from Reinhart *et al.* [1], the latter obtained by stretching nick-labeled DNA for genome mapping in nanochannels at values of α ranging from 0.81 to 8.40 [29]. The results of this analysis are reproduced in Fig. 4.1. The qualitative agreement between theory and these particular experiments is satisfying, in particular the presence of similar fat tails in both theoretical and experimental distributions. However, the agreement between theory and experiment becomes worse as the channel size decreases [128]. This result is counterintuitive; decreasing the channel sizes better satisfies the asymptotic condition $\alpha \gg 1$, which suggests that the agreement between theory and experiment should improve with decreasing channel size. The published analysis [128] only compared theory and experiment [1] for a two distances between nick labels and two channel sizes. We expand this comparison to the entire data set [1] in §4.3; the discrepancy is prevalent throughout all of the data.

One possible source of the discrepancy between theory and experiment is a systematic bias in the experimental data engendered by the genome mapping method used to acquire the data. Explicitly, the DNA in the experimental data set [1] were obtained from *E. coli* cells, labeled in a sequence-specific manner using a nicking enzyme, and then injected into the nanochannel device [29]. In order to identify the genomic distance (in kilobase pairs) between two labels on a given DNA molecule, the label pattern of that molecule needs to be mapped to the reference genome for *E. coli* to identify the

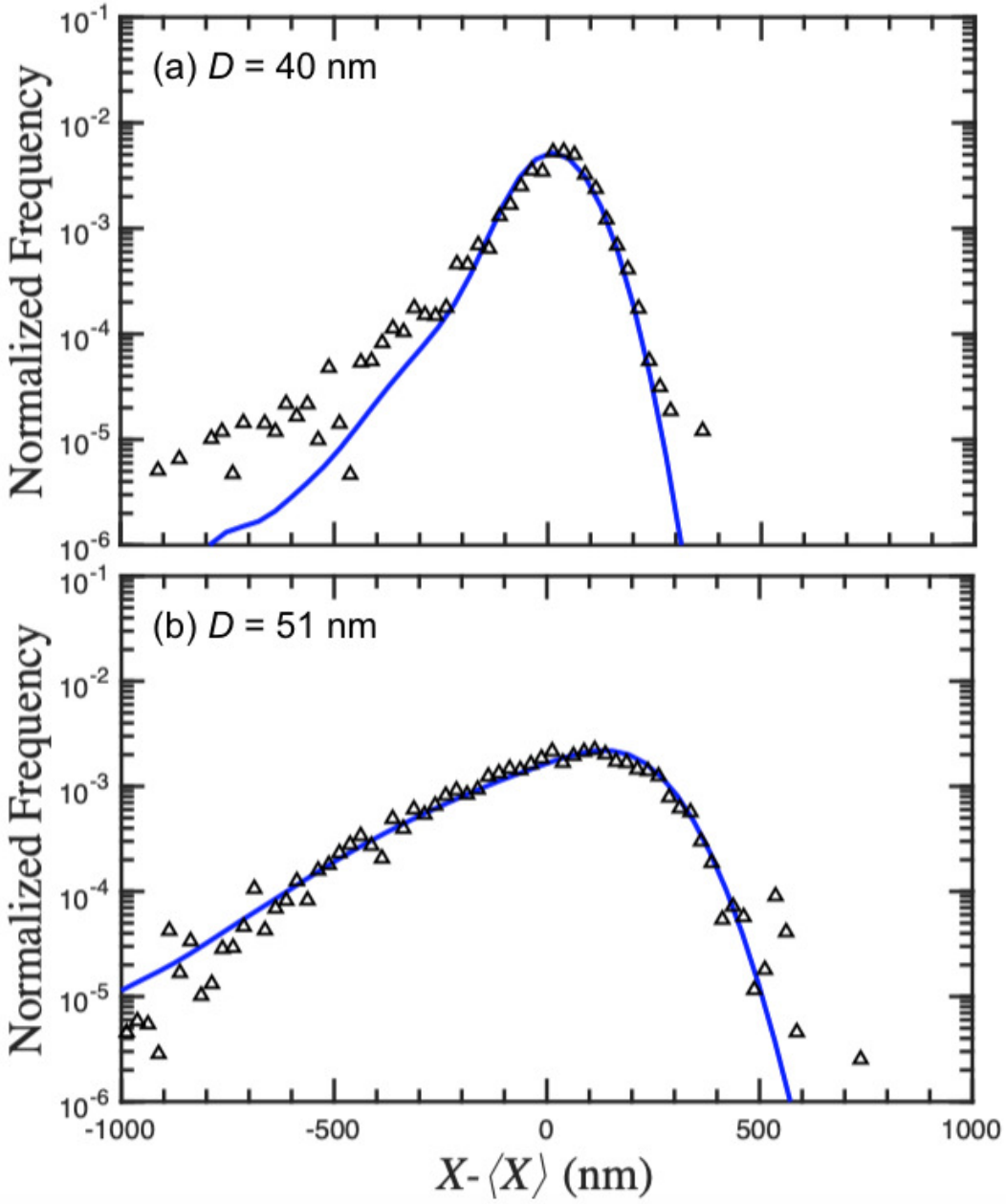

Figure 4.1: Comparison between the predictions of the telegraph model and experimental data in Ref. [1] for the difference between the chain extension, X, relative to the average chain extension, $\langle X \rangle$ for $L = 28{,}125$ bp and two channel sizes: (a) $D = 40$ nm and (b) $D = 51$ nm. Reproduced from Ref. [3].

underlying sequence and, thus, the molecular weight [1]. In our previous experiments, around 25% to 35% of the molecules in a data set that passed the thresholds of length $\geq$ 50 kbp, and number of labels ≥ 5 did not align to the reference [1]. DNA molecules that did not map to the reference were not included in the probability distribution. While many of the unmapped molecules arise from the high variability of the *E. coli* genome, molecules with the correct sequence that cannot be mapped may be due to hairpin folds, which would introduce a systematic bias that cannot be easily resolved using a genome mapping approach. In addition, the spectrum of GC content in the DNA sample may have also confounded the analysis [103], although this is not expected to be a significant factor for reasonably large separations between nick labels [83].

To address these potential concerns about systematic errors in prior experimental work [1], we have undertaken a new experiment described in §4.4 to attenuate all of these experimental artifacts. Explicitly, we repeated the experiments of Reinhart *et al.* [1] using a single DNA sample, λ-phage DNA, without mapping to any reference so as to include all the data points for intact, non-overlapping DNA. Our experiments also probed the extension of DNA at a ratio $D/l_{\mathrm{p}} = 0.62$, and a concomitant value of $\alpha = 88.3$. This ratio of D/l_{p} is even closer to the Odijk regime than previous experiments [1], which only achieved a ratio down to $D/l_{\mathrm{p}} = 0.70$. Comparing the experimental data to the asymptotic solution of the telegraph model [128] reveals that the discrepancy between theory and experiment continues to increase as D/l_{p} decreases, underscoring the conclusions of the previous analysis [128] and ruling out the systematic errors arising in the genome mapping approach [1] on the qualitative conclusions of that analysis. We posit in §4.5 that the long-range effects of DNA-wall electrostatic interactions, which are not included in any of the theories for channel-confined polymers, are the source of the disagreement between the theory and experiment.

4.2 Theory

The one-dimensional telegraph process describes a persistent random walk, where a particle moves with a fixed velocity v_0 on a line for some time T, randomly changing directions with a rate r. The weakly-correlated telegraph process adds a small penalty ε for revisiting the same position along the walk, thereby penalizing frequent changes

in direction. Werner *et al.* [4] showed how the problem of a channel confined DNA molecule can be mapped onto the weakly-correlated telegraph model in the limit of weak excluded volume interactions, i.e. for $D \ll l_{\rm p}^2/w$. The correspondence between the decay in velocity autocorrelation for the telegraph process and the decay in orientational correlations for an confined ideal wormlike chain provides the first step in the mapping, wherein v_0 is replaced by the average orientation a between the ideal polymer backbone and the channel axis and $(2r)^{-1}$ is replaced by the global persistence length g, the latter describing the typical distance between hairpin turns in the absence of excluded volume interactions [129]. Excluded volume appears in the penalty ε for overlapping segments of the confined polymer at a given position within the channel [4]. Importantly, the trio of parameters (a, g, ε) describing the confined polymer problem can all be obtained from simulations of ideal, confined wormlike chains [4]. The total time of the walk, T, corresponds to the contour length of the polymer, L.

Dimensional analysis of the telegraph model reveals that its parameters can only be combined into a single dimensionless number [4]

$$\alpha \equiv \frac{\varepsilon g}{a} \tag{4.1}$$

representing the typical number of overlaps per hairpin turn. The same scaling parameter was obtained by Chen [102] via analysis of the self-consistent field equation description for a confined wormlike chain. The parameters appearing in Eq. (4.1) have been computed by Werner *et al.* [4] over the range of channel sizes relevant to DNA confinement; one can readily map the primitive variables $(l_{\rm p}, w, D)$ describing the confined polymer problem to the telegraph model parameters via interpolation of the data in Ref. [4].

The region of the phase space immediately proximate to the Odijk regime corresponds to small channels, $\alpha \gg 1$, but long chains, $L \gg g$; the cross-over into the Odijk regime takes place when $L \ll g$ and $\alpha \gg 1$, such that there are no hairpin turns in the small channel. In the dual asymptotic limit $\alpha \gg 1$ and $L \gg g$, the telegraph model can be solved [128]. The leading order probability distribution for the chain extension is [128]

$$P(X', T) \sim \mathcal{N} \frac{1 + \sqrt{1 - X'^2}}{(1 - X'^2)^{3/4}} \exp\left[- \left(\frac{L}{2g} \right) S(X') \right], \tag{4.2}$$

where $\mathcal{N}$ is the normalization factor, $X' = X/(aL)$, and the action is

$$S(X') = 3\alpha(1 - X') + 1 - \sqrt{1 - X'^2}. \tag{4.3}$$

The predictions of Eq. (4.2) are in excellent agreement with numerical simulations of the weakly-correlated telegraph process [128], verifying the asymptotic solution of the model.

The telegraph model treats the alignment of the polymer with the channel axis in a mean-field manner through the average alignment a between the polymer backbone and the channel axis, and thus does not include the alignment fluctuations that are present in the Odijk regime [94,95]. As the channel size decreases, the Odijk fluctuations become increasingly significant and eventually overwhelm the rapidly decreasing magnitude of the extension fluctuations due to hairpin formation at large values of α [4]. To account for the Odijk fluctuations, Ödman *et al.* [128] proposed augmenting the telegraph model with the Odijk regime fluctuations via the probability distribution

$$\mathcal{P}(X, L) = \int_0^{aL} dX_1 P(X_1, L)\rho(X_1 - X), \tag{4.4}$$

where

$$\rho(\delta X) = (2\pi\sigma_0^2)^{-1/2} \exp\left[-\frac{\delta X^2}{2\sigma_0^2}\right] \tag{4.5}$$

is a Gaussian with variance σ_0.

In their comparison with experimental data [1], Ödman *et al.* [128] treated σ_0 as an adjustable parameter. In principle, we would anticipate that σ_0 would be the variance in extension in the Odijk regime

$$\sigma^2_{\text{Odijk}} = 0.0096\frac{LD^2}{l_{\text{p}}} \tag{4.6}$$

obtained by Burkhardt *et al.* for a wormlike chain confined to a square channel [95]. We have recently demonstrated [3] that Eqs. (4.4)-(4.6) provide excellent agreement with pruned-enriched Rosenbluth method (PERM) simulations of confined wormlike chains. Indeed, we found that Eq. (4.4) captures simulation data down to $\alpha \approx 0.3$, well below the asymptotic limit $\alpha \gg 1$.

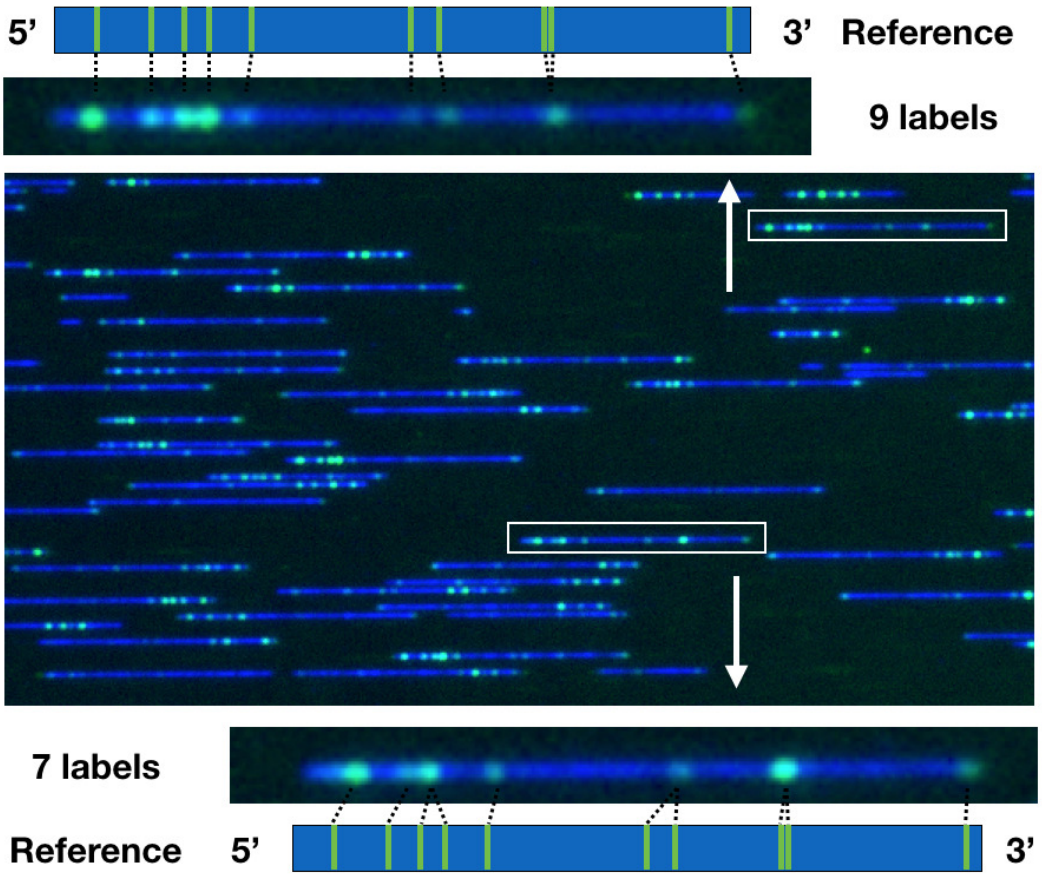

Figure 4.2: Composite, false-color image of λ-DNA molecules with backbone (blue) and the Nt.BspQI nick sites (green) obtained in 34 nm nanochannels. The representative λ-DNA molecules (in the white rectangular boxes) have 9 (upper panel) and 7 (bottom panel) resolvable labels, respectively. The distances between the nearest pairs of labels are measured and compared to the reference. A detailed discussion on filtering data for contour length and the correlation coefficient between the image and the reference is provided in §4.4.1.

4.3 Comparison to Experimental Data for E. coli

We begin our analysis of the telegraph model by comparing its predictions to a set of extension measurement distributions produced by Reinhart *et al.* [1] for the MG1655 *E. coli* strain in five channel sizes, ranging from 40 nm to 51 nm, in a research-grade version of the Bionano Genomics Irys genome mapping system [29]. The image data acquired

in this system are similar to that in Fig. 4.2, which provides a representative image obtained using the approach described in §4.4.1. The channel sizes used by Reinhart *et al.* [1] were close to the DNA persistence length of 52 nm expected for the 0.5x TBE buffer used in those experiments, which has an ionic strength of 103 mM due to residual chemicals used for the nick-labeling reaction that remained in the running buffer [84]. For each molecule, they extracted each section that contained at least 5 contiguous labels that were perfect matches to the reference genome; this region could include the entire molecule. The images provided measurements of the extension X (in nm), while the maps to the reference genome provided the corresponding contour lengths L (in base pairs). A given molecule can contribute many measurements to the total data set, since they contain multiple labels [1, 103]. For comparison to the theory, we assume a conversion of 0.34 nm per base pair due to the low YOYO dye loading (1 dye molecule per 37 bp) used in the experiments. The probability distribution histograms for the experimental extension data used bin sizes of $L = 250$ bp for the separation between nick labels on the DNA to minimize the impact of errors in fluorophore localization on the measured the distance X between the labels. The minimum distance between labels was set to 2500 bp, to remove effects of the diffraction-limited optics on label resolution [130]. Depending on the channel size, these experiments involved between 1839 and 9598 molecules, producing between 8.7×10^4 to 8.0×10^5 measurements of the extension X between label pairs. This data density is sufficient to produce histograms of $X - \langle X \rangle$ for a particular bin in L, where $\langle X \rangle$ is the average extension for label pairs within that bin.

In their comparison of the telegraph model to experimental data, Ödman *et al.* [128] selected two representative values of the contour length between separations, $L = 28{,}125$ bp and 53,125 bp, for both the smallest (40 nm) and largest (51 nm) channel sizes. The variance σ_0 appearing in Eq. (4.5) was treated as an adjustable parameter to fit the right tail of the distribution, with the fitted value of σ_0 being approximately twice that given by Eq. (4.6) [128]. For the effective channel size, Ödman *et al.* considered first the typical approximation [100]

$$D_{\text{eff}} = D - w \tag{4.7}$$

where D is the physical channel size and w is the Stigter effective width [75], which is

D (nm)	D_{eff} (nm)	a	g (nm)	α
40	36.4	0.84	884	8.40
42	38.4	0.84	700	6.15
43	39.7	0.84	656	5.32
49	45.4	0.82	385	2.53
51	47.4	0.81	363	0.81

Table 4.1: Telegraph model parameters for the channel sizes appearing in the experimental data from Reinhart *et al.* [1]. The DNA persistence length is $l_{\text{p}} = 52$ nm and the effective width is $w = 5.6$ nm. The effective channel size is 2 nm greater than that computed from Eq. (4.7) to mimic the best-case scenario of Ödman *et al.* [128] The alignment of the DNA backbone with the channel axis, a, the global persistence length, g, and the scaling parameter, α, were obtained by interpolation to the simulation data of Werner *et al.* [4]

5.6 nm for the experimental conditions [84]. For the experimental data in the 51 nm channel [1], the fitted value of σ_0 led to excellent qualitative agreement between the theory and experiment. However, there was a significant discrepancy in the left tail for the 40 nm channel [128]. which is evident in Fig. 4.1. Ödman *et al.* noted that this discrepancy is attenuated (but not eliminated) by assuming an effective channel size that is 2 nm larger than that computed by Eq. (4.7), and attributed the discrepancy to uncertainties in the actual channel size available to the channel. We will return to this issue in §4.5.

The analysis by Ödman *et al.* [128] suggests there may be a negative correlation for the agreement between theory and channel size, but this tentative conclusion rests on the analysis of only four data points and a qualitative analysis of the data. To provide a firm foundation for their conclusion, we have re-analyzed the entire data set from Reinhart *et al.* [1], and a trio of statistical tests [3] were applied to quantify extent of the discrepancies between theory and experiment for each channel size.

Our analysis follows the approach of Ödman *et al.* [128], using Eq. (4.4) and Eq. (4.5) to model the extension distribution and treating σ_0 in Eq. (4.5) as an adjustable parameter. Table 4.1 provides the relevant experimental parameters and the corresponding telegraph parameter values needed to evaluate Eqs. (4.2)-(4.3). The fitted value of σ_0, required for Eq. (4.4), was determined by fitting the right tail of distribution to the experimental data. These fitting parameters were obtained separately for each channel

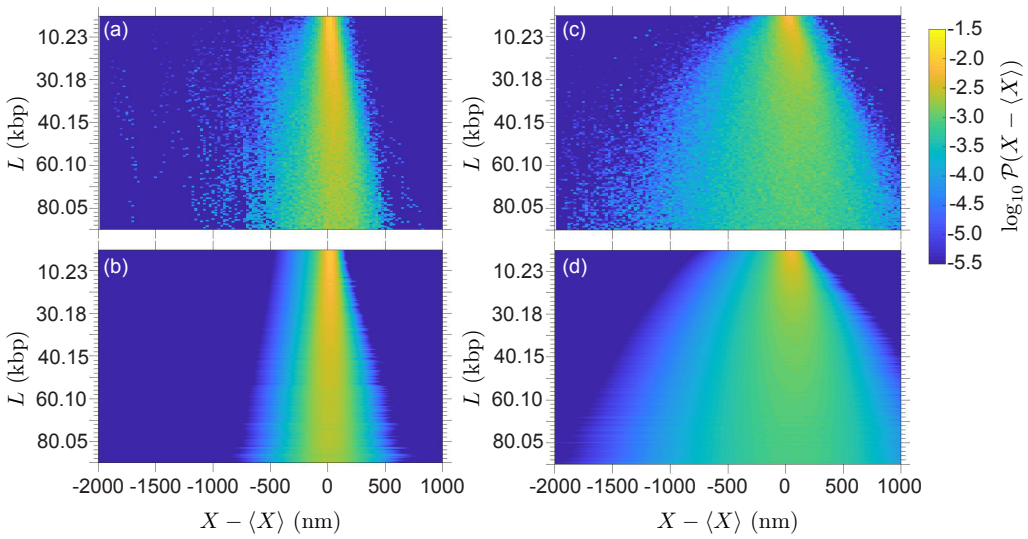

Figure 4.3: Comparison between the predictions of the telegraph model in Eq. (4.4) and the experimental data of Reinhart *et al.* [1] for (a) $D = 40$ nm, experiment; (b) $D = 40$ nm, theory; (c) $D = 51$ nm, experiment; and (d) $D = 51$ nm, theory. Panels (a) and (c) reproduced from Ref. [1].

size D and each bin in L, and appear in Appendix B Fig. B.4. The three smallest channels are fit by $\sigma_0 \approx 1.5\sigma_{\text{Odijk}}$, while the fitting to the two largest channels uses $\sigma_0 \approx 2.5\sigma_{\text{Odijk}}$. Inspired by the prior analysis, we used a 2 nm augmentation to the effective channel size in Eq. (4.7), allowing us to compare the full distributions to the best case scenario in Ref. [128].

Figure 4.3 provides the comparison between theory and experiment [1] over a wide range of contour lengths for the 40 nm channel and 51 nm channel considered previously by Ödman *et al.* [128] Companion figures for the data obtained in the 42 nm, 43 nm and 49 nm channels appear in Appendix B Figs. B.1-B.3. Figure 4.3 confirms that the qualitative conclusions obtained by Ödman *et al.* [128] for two values of L persist for the entire data set. For the 51 nm channel (Fig. 4.3(c) and (d)), the experimental data and theoretical predictions are in excellent agreement, with the main difference being the smoothness of the theoretical result, which is obtained from numerical quadrature, when compared to the experimental data, which have signficant pixelation due to the limited sampling in the tails of the distribution. In contrast, the 40 nm channel data (Fig. 4.3(a) and (b)) exhibit the significant discrepancy in the tails identified previously [128], which

we now see occur for all values of L.

To quantify the degree of agreement between the theoretical and experimental distributions, we used three statistical tests in accordance with our previous work comparing simulations of confined, wormlike chains to the telegraph model [3]. These tests require first converting the probability density functions for each horizontal slice in Fig. 4.3, which corresponds to bin size at a fixed value of L, into cumulative density functions (CDF), which are bounded on the interval $[0, 1]$. The first test computes the root-mean-square (RMS) error between the experimental CDF and the theoretical CDF. The other two tests use an empirical distribution goodness of fit [131] (i) the Cramér-von Mises criterion, which provides uniform weighting to the data and (ii) the Anderson-Darling criterion, which gives more weight in the test to the tails of the distribution. This procedure was repeated for the additional heat maps provided in Appendix B Figs. B.1-B.3.

Figure 4.4 provides the results of the statistical tests as a function of contour length. The absolute values of the deviations measured by these statistical tests increase with the emphasis that the test puts on the tails of the distribution; the RMS error in Fig. 4.4(a) is dominated by the central portion of the data, the Cramér-von Mises criterion in Fig. 4.4(b) incorporates information throughout the distribution with a uniform weight, and the Anderson-Darling criterion in Fig. 4.4(c) puts particular emphasis on the tails of the distribution. The overall trends revealed by these statistics are consistent, with the exception of the anomalously low values in the 42 nm channel. Figure B.1 shows the extension distribution for the 42 nm channels is narrower than anticipated, consistent with previous observations about that particular data set [1].

All of the data exhibit an initial decrease in the deviation with increasing contour length. We attribute this trend to the asymptotic nature of the solution to the telegraph model; the chain needs to be sufficiently long to form hairpin bends ($L \gg g$), which is not satisfied for the shortest contour lengths. In general, this initial transient in the deviation as a function of contour length decays at a label separation of approximately 20 kbp, supporting the smaller value of the contour length selected by Ödman *et al.* [128].

The main goal of our re-analysis of the experimental data [1] with the telegraph model [128] is to determine how the deviation between the theory and experiment depends on the channel size. Figure 4.5 provides the results obtained using each metric (RMS error, Cramér-von Mises, and Anderson-Darling) as a function of channel size,

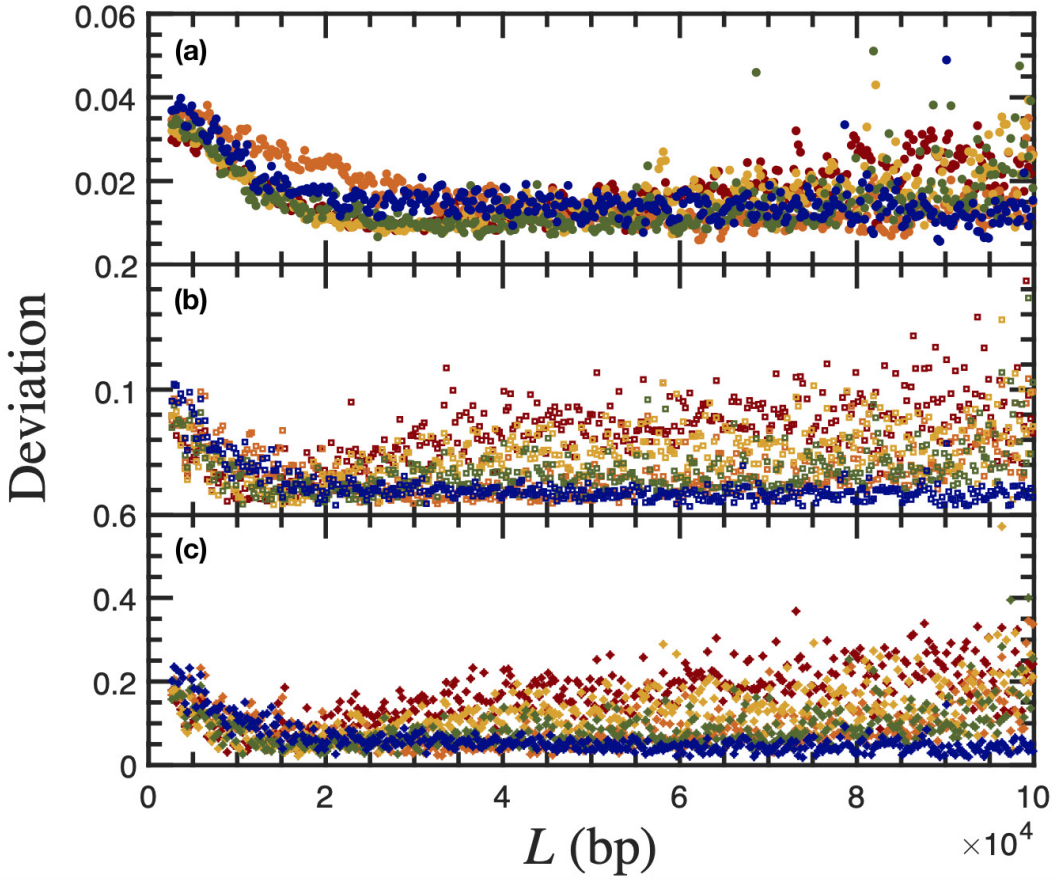

Figure 4.4: The result of statistical tests for quantifying the degree of agreement between theory and experiments: (a) the RMS error, (b) the Cramér-von Mises criterion, and (c) the Anderson-Darling criterion. Statistical data were obtained for 40 nm (red), 42 nm (orange), 43 nm (yellow), 49 nm (green), and 51 nm (blue) nanochannels using the probability distributions in Fig. 4.3 and Figs. B.1-B.3 in Appendix B.

where we have averaged the results in Fig. 4.4 from $L = 20$ kbp to $L = 80$ kbp to remove the aforementioned artifacts that arise at small and large molecular weights. The trends in Fig. 4.5 are consistent with the conclusions drawn by Ödman *et al.* [128] from their analysis of a limited data set: the deviation between experiment and theory increases with decreasing channel size, with the exception of the 42 nm channels. As noted in the context of Fig. 4.4, the magnitude of the deviations for the different statistical tests reflects their emphasis on the tails of the distribution, which is strongest for the

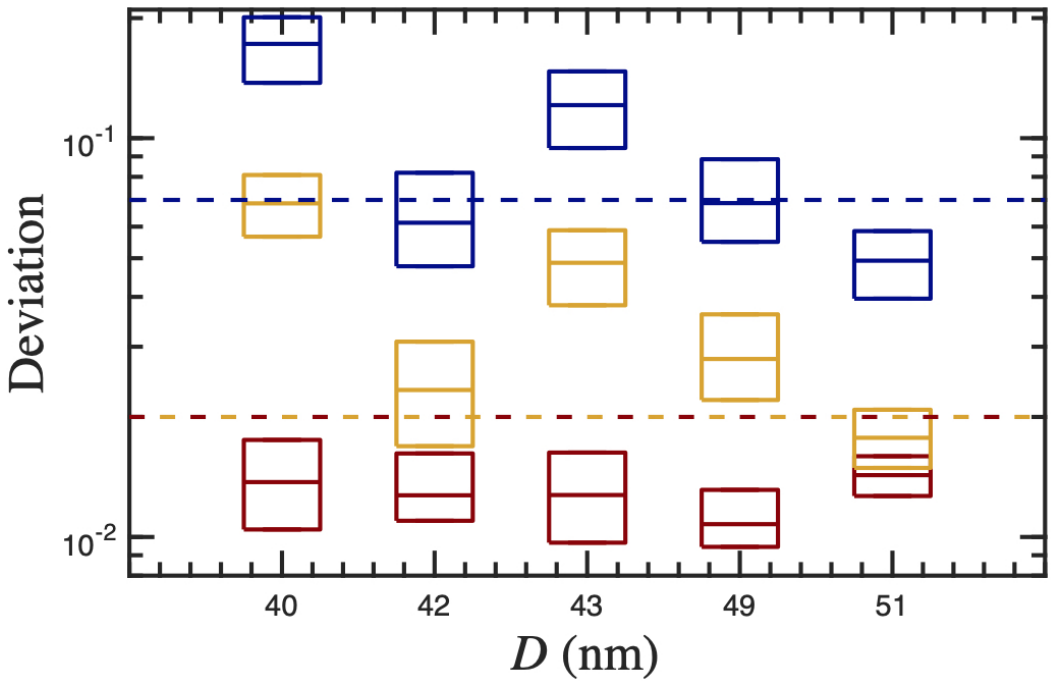

Figure 4.5: The result of statistical tests as a function of channel size by averaging the results in Fig. 4.4 from $L = 20$ kbp to $L = 80$ kbp for quantifying the degree of agreement between theory and experiments: the RMS error (red), the Cramér-von Mises criterion (yellow), and the Anderson-Darling criterion (blue). The boxes span from 25% to 75% of the data, with the lines indicating the median value. For the comparison between theory and simulation, the dashed line alternating between red and yellow indicates the value of 0.02 which was obtained by the RMS error and Cramér-von Mises criterion; and the blue dashed line is at the value of 0.07 obtained by the Anderson-Darling criterion [3].

Anderson-Darling test and weakest for the RMS error.

It can be challenging to readily interpret the magnitude of the deviations in Fig. 4.5 without a familiarity with these types of measurements. By way of comparison, we recall that the plots of the CDFs obtained from our recent simulations of confined wormlike chains [3] were almost indistinguishable from the telegraph model CDFs for large values of α. The corresponding values of the statistical criteria for $\alpha > 1$ were approximately 0.07 for the Anderson-Darling criterion (blue dashed line in Fig. 4.5), and 0.02 for both the RMS error and the Cramér-von Mises criterion (dashed line alternating between red and yellow in Fig. 4.5). Figure 4.5 indicates that the deviations for the wider

channels (49 nm and 51 nm) are similar to those obtained for comparisons between theory and simulation [3], but the deviations from the Cramér-von Mises criterion and the Anderson-Darling criterion, are substantially higher for experiments than simulation as the channel size decreases. The RMS error deviation remains roughly constant with channel size, indicating that the theory does an adequate job in capturing the central portion of the probability distribution, which dominates the RMS error.

It is important to note that the comparison between experiment and the telegraph theory leading to Fig. 4.5 allows σ_0 to be a fitting parameter. Had we required the theoretical calculations to use the Odijk theory result in Eq. (4.6), which was the case in the comparison between simulation and theory [3], the deviations between experiment and theory would increase further.

4.4 Comparison to Experimental Data for λ-DNA

Our re-analysis of the experimental data from Reinhart *et al.* [1], most notably the statistical tests in Fig. 4.5, suggest an increasing discrepancy between theory and experiment as the channel size decreases. However, one concern about the genome mapping approach used previously [1] is that it requires mapping the labeled DNA back to the reference genome to obtain their contour length. Some molecules do not align to the reference genome, around 25% to 35% of the molecules in a data set. In cases where the lack of alignment is due to incomplete nick labeling or changes in the host genome, excluding these molecules does not impact our physical measurements. However, molecules that do not align for physical reasons, in particular significant backfolding, are excluded from the final data set. This is unlikely to be the source of the discrepancy, since we would expect the exclusion of such molecules to lead to the theory overestimating the left tail of the extension distribution; this is opposite the behavior shown by Ödman *et al.* [128] and the results in Fig. 4.3(a) and (b).

To rule out the confounding effects due to mapping to the reference genome to obtain the value of L, this section describes a new experiment using λ-DNA (48.5 kbp) as the model polymer. These experiments also took advantage of (i) an experimental protocol that removes the labeling chemicals prior to injection into the nanochannel [46, 47], allowing us to lower the ionic strength to 48 mM and a corresponding persistence length

[67] of $l_p = 54.9$ nm, and (ii) the Bionano Genomics Saphyr chip, which lowered the channel size to $D = 34$ nm while retaining the high-throughput required to obtain the tails of the extension probability densities [1]. The latter channel size is similar to the 30 nm $\times$ 40 nm lower bound in the classic experiment by Reisner *et al.* [96], thereby probing a very low value $D/l_p = 0.62$. If we adopt Eq. (4.7) as the approximation for the effective channel size and use the DNA effective width $w = 7.6$ nm predicted by Stigter's theory [75], the effective channel size is $D_{eff} = 26.4$ nm and the ratio $D_{eff}/l_p = 0.48$ is considerably smaller than unity. We will examine the validity of this effective channel size in §4.5.

4.4.1 Experimental Method

λ-DNA (48.5 kbp, New England Biolabs) was labeled at the 5'-GCTCTTC-3 site using the nicking, labeling, repairing, and staining (NLRS) protocol [46, 47]. Briefly, the λ-DNA were nicked with the nicking enzyme Nt.BspQI, labeled by inserting a cy3-like fluorophore during the labeling step, repaired using *Taq* ligase, and stained by YOYO-1 at a dye to base pair ratio of 1:37. Drop dialysis with Tris-EDTA (TE) buffer was applied to remove the extra reagents in the previous steps and the DNA was suspended in the Bionano Genomics running buffer, which has an ionic strength of 48 mM. The experiment was conducted with a BioNano Genomics Saphyr chip, which contains an array of 34 nm wide nanochannels. The uncertainty in the nanochannel width is estimated as ± 2.5 nm [83]. The labeled molecules were driven into nanochannels by an electrokinetic system, and then sequentially excited by a sapphire green laser (532 nm, 300mW, Coherent) and an OBIS blue laser (488 nm, 300mW, Coherent) to image the DNA backbone and nick labels, respectively, on a research-grade version of the Bionano Genomics Saphyr system. Figure 4.2 shows a typical false-color image from the experiment, with blue and green channels combined.

The acquired image stack includes a total of 1,241,907 molecules. The images were analyzed by Bionano Genomics's image processing algorithm (available from Bionano Genomics) to identify the DNA molecules, measure the extension of their backbone (blue laser) along the channel axis, and measure the distance between nick sites (green laser). Owing to its use for genome mapping, the image processing algorithm reports its output in units of base pairs. The program output was converted to physical distances

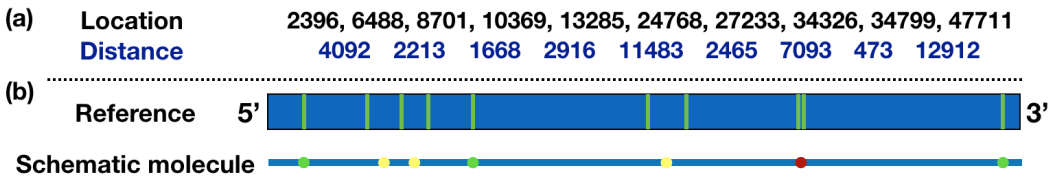

Figure 4.6: Scheme of λ-DNA reference with (a) the locations of all the nick sites counted from 5' end and the distance between every nearest pairs of nick sites, and (b) a schematic molecule that illustrates the diffraction-limited spots. Due to the diffraction-limited optics, proximate nick sites can merge into a single label if the distance between pairs of nick sites is less than 1,500 to 2,500 bp. The pair of nick sites that always appear as a single label are labeled in red (the case of 8^{th} and 9^{th} nick sites); the nick sites that may be a single label due to fluctuations are labeled in yellow; all other resolvable nick sites are labeled in green.

using the conversions 366 bp = 1 pixel for the algorithm, which is a built-in conversion factor in the code, and 1 pixel = 0.1083 μm for the microscope optics.

There exist a number of experimental challenges when using the high-throughput data obtained by this method for physical measurements, such as incomplete dye labeling, shear fragmentation and photocleavage of the DNA, and overlapping molecules, since the images are acquired without the typical low-throughput approach of screening the field of view to confirm that the molecules appear to be intact and labeled. These are not significant issues for the practice of genome mapping [29], since producing a consensus genome map essentially averages over the data. Since we are interested in analyzing the tails of the extension distribution, which involve the rare events, we need to ensure that we only analyzed intact, isolated λ-DNA molecules. To this end, we adopted the conservative approach of analyzing the extension distribution between the most distant pair of nick sites of λ-DNA, rather than the extension of the entire molecule, to ensure that the region between these nick sites (45,315 bp far apart) was intact. We also adopted a set of three filters that attenuate the effects of systematic errors in the data set.

In the first filtering step, we confirmed that the molecule has the correct number of labels. The λ-DNA genome has ten 5'-GCTCTTC-3' nick sites, illustrated in Fig. 4.6. However, proximate nick sites can merge into a single, diffraction-limited label if the distance between pairs of nick sites is less than 1,500 to 2,500 base pairs. This is always

the case for the 8^{th} and 9^{th} nick sites, which are only separated by 473 bp. It is also possible that small hairpins can lead to a collapse of the other label pairs indicated in Fig. 4.6. As a consequence, an image of a completely labeled molecules will have between 7 and 9 resolvable labels. Fig. 4.2 shows two representative molecules with 9 (upper panel) and 7 (bottom panel) resolvable labels, respectively. Requiring that all molecules meet this condition reduced the data set to 612,121 molecules.

In the second filtering step, we required that molecules have a backbone extension L between 14.0 μm and 16.5 μm. The latter value corresponds to fully extended λ-DNA. The former value corresponds to the distance between the first and last nick points if the molecule were stretched to the 91.2% extent predicted by Odijk's theory [93, 95], which is a robust lower bound since there is additional DNA on both sides of those nick sites. To ensure that our key conclusions are not affected by the choices for these upper and lower bounds, the Appendix B provides a sensitivity analysis to the parameter selection. This additional length filter reduced the total size of the data set to 509,070 molecules.

Additional experimental artifacts can arise from small DNA fragments that become adsorbed to the walls and are imaged along with the free molecules, proximate DNA fragments within the channel that are erroneously identified as a single molecule [86], and free dye molecules [85]. To attenuate these effects, the third filtering step checked that the labeling pattern on the DNA was consistent with the expected pattern for λ-DNA. While the scoring methods used in genome mapping could be used here, analogous to previous work [86, 103], we chose to adopt a different approach that was tailored to the particular details of the reference λ-DNA pattern in Fig. 4.6 that can accommodate readily the diffraction-limited spots. For molecules with 9 labels, we computed the correlation coefficient between the measured distance between labels and the reference. For molecules with 8 labels, the same calculation was performed for all possible 8-label combinations of Fig. 4.6; the calculation for molecules with 7 labels used all possible 7-label combinations. In all cases, we considered both the patterns produced by Fig. 4.6 and their inverse, since the 5'-3' orientation of the DNA within the nanochannel is not known. For completeness, all of these possible patterns are included in the Fig. B.5 of Appendix B. We then selected the highest correlation obtained from all of the calculations for that molecule. Only those molecules whose largest correlation coefficient exceeded a cutoff of 0.98 were included for further analysis. This value was selected

D_{eff} (nm)	a	g (nm)	α	σ_{Odijk} (nm)
26.4	0.88	4436	88.3	43.2
34.0	0.85	1289	16.3	55.7
36.5	0.85	973	10.6	59.8

Table 4.2: Telegraph model parameters for the channel sizes appearing in the experimental data obtained from λ-DNA in a $D = 34$ nm channel. The DNA persistence length is $l_{\text{p}} = 54.9$ nm and the effective width is $w = 7.6$ nm. The effective channel sizes considered are the estimated physical channel size (34 nm), the upper bound in the physical channel size (36.5 nm), and the approximation in Eq. (4.7). The alignment of the DNA backbone with the channel axis, a, the global persistence length, g, and the scaling parameter, α, were obtained by interpolation to the data of Werner *et al.* [4]. The values of σ_{Odijk} are calculated using Eq. (4.6). The value $\sigma_0 = 199$ nm is the best-fit parameter to capture the right tail of the distribution and independent of the choice of D_{eff}.

by first simulating 10^4 patterns by randomly selecting locations within the simulated molecule to insert a hairpin using the frequency and hairpin size predicted by Ödman *et al.* using the telegraph model [128]. Simulated label locations below 1,500 bp were assumed to merge into a single label, and simulated label locations between 1,500 and 2,500 bp were merged into single labels by selecting a uniformly distributed random variable between those two limits for each molecule and merging an labels that were separated less than that value. The correlation coefficients for these simulated data with the reference produced the lowest value of 0.98. This final filter reduced the data set to 166,340 molecules for analysis. Compared to the molecules left from the previous filtering step, around 67% molecules were removed in the last correlation coefficient filter. More data points could be included by lowering the cutoff of correlation coefficient, and the corresponding sensitivity analysis to the choice of the cutoff value is provided in Appendix B. Inasmuch as the result from the sensitivity analysis in Appendix B shows that the adjustment would only broaden the extension distribution and make our conclusion more significant, here we used the final data set, 166,340 molecules, with the strictest cutoff setting for the further discussion.

	$\sigma_0 = \sigma_{\text{Odijk}}$			$\sigma_0 = 2\sigma_{\text{Odijk}}$			Best fit σ_0		
D_{eff} (nm)	26.4	34.0	36.5	26.4	34.0	36.5	26.4	34.0	36.5
RMS error	0.10	0.093	0.090	0.083	0.069	0.065	0.040	0.040	0.040
Cramér-von Mises	0.22	0.19	0.198	0.16	0.12	0.11	0.039	0.037	0.036
Anderson-Darling	1.33	1.00	0.90	0.72	0.48	0.41	0.12	0.12	0.11

Table 4.3: Results of the statistical tests for the data in Fig. 4.7 and Fig. B.8 in Appendix B.

4.4.2 Results

The comparison of the experimental data with the telegraph model requires selecting an appropriate channel size D_{eff}. We have considered the three possible cases listed in Table 4.2. The typical approximation [100] given by Eq. (4.7) furnishes $D_{\text{eff}} = 26.4$ nm for a physical channel size $D = 34$ nm and the effective width $w = 7.6$ nm at this ionic strength. We also considered the case where we neglect the long-ranged DNA-wall excluded volume interactions using the most likely channel size, corresponding to $D_{\text{eff}} = 34$ nm, and the same assumption using the upper bound in the estimate for the channel size [83], $D_{\text{eff}} = 36.5$ nm. We chose to examine the latter value to provide the most favorable possible comparison between the theory and experiment, as we know from prior work [128] that the theory comes closer to the experimental data [1] by increasing D_{eff}. Nevertheless, if the overall modeling approach is applicable, we would expect the channel size $D_{\text{eff}} = 26.4$ nm to be the best estimate based on the success of this approach in modeling DNA confinement in larger channels [80–82, 100].

Figure 4.7 compares the predictions of the telegraph model to the experimental data for each of the values of D_{eff} in Table 4.2, using the telegraph model parameters in Table 4.2. The dotted lines treat the alignment fluctuations using the Odijk variance σ_{Odijk} in Eq. (4.6), while the solid lines use the best-fit value $\sigma_0 = 199$ following the approach used by Ödman *et al.* [128] and our own analysis in §4.3. We also provide a comparison between experiment and theory using the correction $\sigma_0 = 2\sigma_{\text{Odijk}}$ proposed by Ödman *et al.* [128] in Appendix B Fig. B.8. Table 4.3 provides the statistical tests for comparison to the λ-DNA experiments.

The disagreement between theory and experiment in the λ-DNA experiments is even more prominent for this small channel than we observed for the *E. coli* sample in

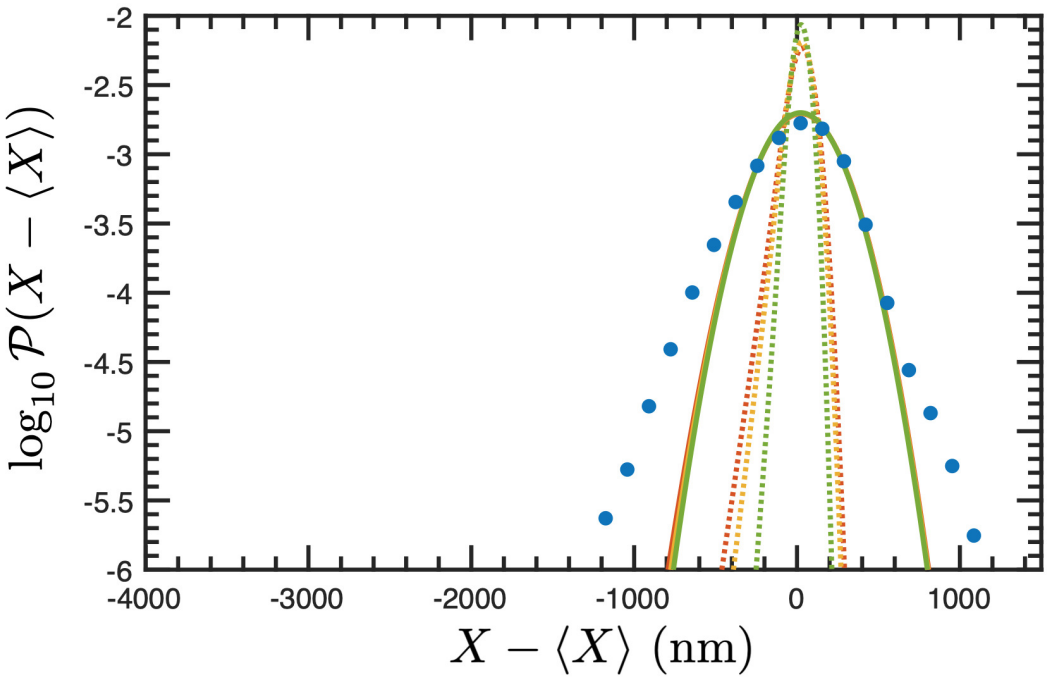

Figure 4.7: Comparison between the predictions of telegraph model with $\sigma_0 = \sigma_{\text{Odijk}}$ (dotted lines) and the best fit for σ (solid lines) and the experimental data of λ-DNA (blue circle). The theoretical distributions were model at $D_{\text{eff}} = 36.5$ nm (red), $D_{\text{eff}} = 34$ nm (yellow), and $D_{\text{eff}} = 26.4$ nm (green) nanochannels using the parameter values in Table 4.2.

the larger channels. These results indicate that the trend of increasing deviation with decreasing channel size is robust to the methodology, and not an artifact of the genome mapping approach used previously [1]. The shortcomings in the model are especially apparent when the Odijk variance is used to model the alignment fluctuations; the theory describes neither the right nor left tails.

Allowing the variance σ_0 to be an adjustable parameter alleviates the problem somewhat by broadening the predicted distribution to better reflect the experimental results. However, there are two notable issues with this semi-empirical approach. First, treating σ_0 as a fitting parameter becomes less effective as the channel size decreases. We have already noted this effect in the context of Fig. 4.5, with the statistical measures of the deviation increasing with decreasing channel size. Comparing those results for *E. coli* DNA to Table 4.3 reveals a continued increase in the deviation between experiments

and theory for the 34 nm channel, even when we make the questionable approximation of allowing the DNA to occupy the entire channel cross-section. Second, the deviation between the best-fit value of σ_0 and that predicted for the Odijk regime [95] is increasing with decreasing channel size. The best-fit values of σ_0 used by Ödman *et al.* [128] were typically a factor of two larger than the Odijk theory for their limited analysis, and our analysis of the entire data set (see Appendix B Fig. B.4) suggests this is a reasonable approximation. In contrast, Table 4.3 indicates that capturing qualitatively the experimental distribution for $D = 34$ nm required using a five-fold increase in the variance with respect to the Odijk theory. While combining the telegraph model with a fitted variance due to alignment fluctuations may provide a functional description of the experimental data, the inability to predict the required value of σ_0 — and the significant and growing deviations between the best-fit value and the Odijk theory with decreasing channel sizes — limits the predictive value of the model.

A possible source of the discrepancy between the theory and experiment lies in the parameters used to filter the outliers (incomplete labeling, fragmented molecules) in the data set. We have taken a rather conservative approach to the filtering, as indicated by the attrition between the original data set and those molecules that were deemed to be intact, λ-DNA. Appendix B provides a sensitivity analysis to the parameter selection of the cutoff values for the length filter and the correlation coefficient filter. This additional analysis indicates that the disagreement between the theory and the experiment was unaffected by having a stricter restriction on the length filter. The disagreement becomes more prominent when lowering the cutoff values of correlation coefficient, which broadened the experimental distribution by including molecules that were previously excluded due to poor alignment. We thus conclude that our analysis is robust to the particular parameters choices used to filter the data.

4.5 Discussion

It is clear from the analysis of our experimental data for *E. coli* DNA [1] and λ-DNA that the agreement between the predictions of the telegraph model and the data becomes worse as the channel size decreases. The salient question is then whether the problem lies in the telegraph model itself, or the applicability of this particular model to the

experimental data. Since we have shown elsewhere [3] that the predictions of Eq. (4.4) are in very close agreement with simulation data for a discrete wormlike chain model, it appears that the problem lies in the applicability of the this physical model to describe the experimental scenario.

The most likely shortcoming in the models used to describe DNA in nanochannel confinement is that these models [4,48,49,102] are derived for a neutral polymer confined between hard walls. Using such models to interpret experimental data for DNA is predicated on the assumption that the confined polyelectrolyte physics can be mapped to an equivalent neutral polymer. The persistence length l_p [52,53,67] and DNA effective width w [75] are mapped using polyelectrolyte theory for unconfined polymers. The challenge lies in the DNA-wall interactions, which are conventionally handled by defining an effective channel size $D_\mathrm{eff} < D$ that accounts for the repulsion of the DNA from the walls [44, 100].

Let us consider, in the context of a relatively simple model, the extent to which DNA-wall electrostatic interactions are expected to affect the extension of DNA in a nanochannel. In this context, it is convenient to consider what happens in the Odijk regime, where excluded volume effects are negligible. Following the notation of Chen [132], the Hamiltonian for a wormlike chain in an external potential is

$$\frac{\mathcal{H}}{k_\mathrm{B}T} = \int_0^1 dt \left\{ \frac{l_\mathrm{p}}{2L} \left| \frac{\mathrm{d}\mathbf{u}(t)}{\mathrm{d}t} \right|^2 + \frac{U[\mathbf{r}(t), \mathbf{u}(t)]}{k_\mathrm{B}T} \right\}, \tag{4.8}$$

where $\mathbf{r}$ is the spatial position within the channel, $\mathbf{u}$ is the (dimensionless) tangent vector to the backbone chain, $t \in [0, 1]$ is the fractional distance along the polymer backbone, and $k_\mathrm{B}T$ is the Boltzmann factor. The potential $U[\mathbf{r}(t), \mathbf{u}(t)]$ is defined per unit length, such that integration over t provides the total external energy for the chain. The propagator $q(\mathbf{r}, \mathbf{u}; t)$ corresponding to Eq. (4.8) satisfies the modified diffusion equation [132]

$$\frac{\partial}{\partial t} q(\mathbf{r}, \mathbf{u}; t) = \left[-L\mathbf{u} \cdot \nabla_\mathbf{r} + \frac{L}{2l_\mathrm{p}} \nabla^2_\mathbf{u} - \frac{U(\mathbf{r}, \mathbf{u})}{k_\mathrm{B}T} \right] q(\mathbf{r}, \mathbf{u}; t). \tag{4.9}$$

The Odijk theory can be derived [94] from the case of a neutral polymer interacting

with hard walls, where the propagator satisfies

$$\frac{\partial}{\partial t}q(\mathbf{r},\mathbf{u};t) = \left[-L\mathbf{u}\cdot\nabla_{\mathbf{r}} + \frac{L}{2l_{\mathrm{p}}}\nabla^2_{\mathbf{u}}\right] q(\mathbf{r},\mathbf{u};t), \tag{4.10}$$

subject to appropriate boundary conditions on the hard walls [132]. At the scaling level [94, 133], using $\mathbf{r} \sim D$ and balancing the two terms in the operator of Eq. (4.10) reveals that $\mathbf{u} \sim (D/l_{\mathrm{p}})^{1/3}$. The time in the propagator is then $t \sim \lambda/L$, where

$$\lambda \equiv \left(l_{\mathrm{p}}D^2\right)^{1/3} \tag{4.11}$$

is the Odijk deflection length [93]. Determining the extension of the chain requires computing the thermal fluctuations for an almost completely stretched chain [93, 95]

$$X = L\left[1 - \frac{1}{2}\int_0^1 d\mathrm{t}\langle\mathbf{u}(t)^2\rangle\right]. \tag{4.12}$$

The scaling $\langle\mathbf{u}^2\rangle \sim (D/l_{\mathrm{p}})^{2/3}$ leads to the well known result [93]

$$X = L\left[1 - A\left(\frac{D}{l_{\mathrm{p}}}\right)^{2/3}\right] \tag{4.13}$$

The prefactor A can be obtained by solving Eq. (4.10) using ground-state dominance and appropriate boundary conditions for a hard wall [132, 133]. Alternatively, one can convert the problem into an accelerated particle model [94, 95] or simulate an ideal wormlike chain using a particle-based model [134]. These different approaches all yield very similar results [132]; we have typically invoked the result $A = 0.18274$ obtained by Burkhardt *et al.* [95] for square channels.

Now let us consider the case where there is a potential that governs the polymer-wall interactions, in excess of the infinite energy penalty associated with penetrating a hard wall. It proves convenient here to recast Eq. (4.9) into a dimensionless form. Following the preceding analysis, we define a dimensionless position $\tilde{\mathbf{r}} = \mathbf{r}/D$, scaled tangent vector $\tilde{\mathbf{u}} = \mathbf{u}(l_{\mathrm{p}}/D)^{1/3}$, and dimensionless time $\tilde{t} = tL/\lambda$. Let us further write the dimensionless interaction potential $\phi(\mathbf{r},\mathbf{u}) = U(\mathbf{r},\mathbf{u})/U_0$, where U_0 is the maximum

value of U. The dimensionless form of Eq. (4.9) is

$$\frac{\partial}{\partial \tilde{t}} q(\tilde{\mathbf{r}}, \tilde{\mathbf{u}}; \tilde{t}) = \left[-\tilde{\mathbf{u}} \cdot \nabla_{\tilde{\mathbf{r}}} + \frac{1}{2} \nabla^2_{\tilde{\mathbf{u}}} - \beta \phi(\tilde{\mathbf{r}}, \tilde{\mathbf{u}}) \right] q(\tilde{\mathbf{r}}, \tilde{\mathbf{u}}; \tilde{t}) \tag{4.14}$$

where

$$\beta \equiv \frac{U_0(\lambda/L)}{k_{\mathrm{B}}T} \tag{4.15}$$

is the polymer-wall interaction energy per deflection segment.

We would expect to recover the Odijk statistics from Eq. (4.14) in the limit $\beta \ll 1$, independent of the particular form of $\phi(\mathbf{r}, \mathbf{u})$. In this case, the polymer-wall interactions would represent a small correction, on par with other approximations in the Odijk theory [93]. We would also expect to recover Eq. (4.13) for cases where the function $\phi(\mathbf{r}, \mathbf{u})$ attenuates sufficiently fast such that $\beta\phi(\mathbf{r}, \mathbf{u})$ is sensible only in a thin boundary layer near the wall. In this case, it is convenient to define an effective channel size similar to Eq. (4.7) that approximates the exclusion of the DNA from the region immediately proximate to the wall, which is the standard approach for accounting for DNA-wall interactions when analyzing experimental data [44, 83, 100].

For our purposes, we need to determine whether $\beta\phi(\mathbf{r}, \mathbf{u})$ is sufficiently attenuated at the experimental conditions used here ($D = 34$ nm, $I = 48$ mM). From Dobrynin's theory [67], the persistence length is $l_{\mathrm{p}} = 55$ nm and the corresponding deflection segment length from Eq. (4.11) is $\lambda = 40$ nm. For the DNA-wall interaction energy, we adopt the simple model of a DNA segment of effective charge density ν_{eff} moving in an electrostatic potential $\psi(z) = \Psi_{\mathrm{wall}} e^{-\kappa z}$ created by the channel walls [44]

$$\frac{U}{L} = \nu_{\mathrm{eff}} \Psi_{\mathrm{wall}} e^{-\kappa z} \tag{4.16}$$

where z is the distance measured perpendicular to the surface of the wall. In the latter, the Debye length is

$$\kappa^{-1} = \sqrt{\frac{\epsilon_0 \epsilon_b k_{\mathrm{B}} T}{2 N_A e^2 I}} \tag{4.17}$$

with ϵ_0 the permittivity of free space, $\epsilon_b = 80$ the dielectric constant of the buffer, N_A is Avogadro's number, and I the ionic strength of the buffer. For our conditions, the Debye length is $\kappa^{-1} = 1.4$ nm. This model assumes, *inter alia*, that the electrostatic

problem is one-dimensional. If the Odijk theory can be applied, this assumption will be reasonable since the DNA-wall interactions will be sensible only within a thin layer near the wall, which can be approximated as a flat interface without a significant contribution from the corners.

Both the effective charge of the DNA and the surface potential are affected by the ionic strength of the buffer. For ν_{eff}, we use the approach Stigter [75] developed in the context of computing the effective width of DNA,

$$w = \frac{1}{\kappa}\left[0.7704 + \ln\left(\frac{\nu_{\text{eff}}^2}{2\epsilon_0\epsilon_b k_{\text{B}}T\kappa}\right)\right] \tag{4.18}$$

For a given ionic strength, ν_{eff} is computed numerically by matching the far-field solution of the Guoy-Chapman model for a charged cylinder of charge density ν to the far-field solution of the Debye-Hückel model for a charged cylinder of charge density ν_{eff} [135]. Numerical data are available [64] for w as a function of ionic strength, which can be readily inverted into data for ν_{eff} via Eq. (4.18). For our experimental conditions, ν_{eff} = 4.2 e/nm.

For Ψ_{wall}, we need to compute the ζ potential of the surface [44],

$$\Psi_{\text{wall}} = \frac{4k_{\text{B}}T}{e}\tanh\left(\frac{e\zeta}{4k_{\text{B}}T}\right) \tag{4.19}$$

The ζ potential for fused silica can be obtained from the model of Behrens and Grier [136], which involves the simultaneous solution for the ζ potential,

$$\frac{\zeta e}{k_{\text{B}}T} = \ln\left(\frac{-\sigma}{e\Gamma + \sigma}\right) - (\text{pH} - \text{pK})\ln(10) - \frac{\sigma e}{Ck_{\text{B}}T} \tag{4.20}$$

and the surface charge density,

$$\sigma = \frac{2\epsilon\epsilon_0 k_{\text{B}}T\kappa}{e}\sinh\left(\frac{e\zeta}{2k_{\text{B}}T}\right) \tag{4.21}$$

where pH = 8.6 is a the typical buffer condition [83], and, for silica and glass, pK = 7.5 is the dissociation constant [137], C = 2.9 F/m^2 is the Stern layer capacitance [137], and Γ = 8 nm^{-2} is the surface density of ionizable groups [138]. The resulting wall potential is -77 mV.

Combining the estimates for λ and $U_0/L = \nu_{\rm eff}\Psi_{\rm wall}$ leads to $\beta = 500$. Clearly, one cannot completely ignore the wall-interaction term in Eq. (4.9). This conclusion agrees with intuition. DNA is a highly charged object and, in these buffer conditions, silica is also highly charged and repulsive for the DNA; any attempt to bring the DNA into contact with the wall is going to be strongly repelled.

The more relevant question is the length scale for the DNA-wall interactions. If the DNA-wall effect is short-ranged, then one still might expect to observe Odijk statistics for DNA in a nanochannel, albeit with an effective channel size that is somewhat smaller than the actual channel due to the local repulsion. Unfortunately, we find that the magnitude of the wall interaction term, $\beta\phi$, only decays to the relatively small value $\beta\phi \approx 0.1$ after 8.5 Debye lengths. For our experiments, this decay length corresponds to $z^* = 12$ nm away from the wall, a substantial fraction of the 17 nm half-width of the channel. It seems implausible that such a long-ranged interaction can be accurately captured by defining an effective channel size $D_{\rm eff} = 10$ nm.

It is worthwhile to consider whether the conclusions drawn for our particular experimental system ($D = 34$ nm, $I = 48$ nm) hold for other experimental systems that are proximate to the Odijk regime. Figure 4.8 provides data for the dimensionless wall interaction parameter β and the position z^* where the wall interaction energy $\beta\phi(z)$ decays to 0.1 $k_{\rm B}T$. The solid black curve extends the results for our particular channel size to a wide range of ionic strengths. Even at the relatively high ionic strength of 100 mM, the wall interaction remains sensible out to a considerable distance $z^*/D = 0.25$. As the system probes deeper into the Odijk regime ($D/l_{\rm p} = 0.1$), the effect of wall interactions only becomes more important. We have also considered the marginal case of $D = l_{\rm p}$, where our analysis becomes suspect due to the emergence of excluded volume interactions. Neglecting that complication, we again see that DNA-wall electrostatic interactions remain very prominent at lower ionic strengths. Even at an ionic strength of 100 mM, the effect of a single wall continues to persist to almost 20% of the channel cross section. We thus conclude that DNA-wall electrostatic interactions play an important role for all experimentally relevant scenarios for confinement in channels near the persistence length.

Clearly, the theoretical analysis pursued here is a simplified description of the experimental scenario, neglecting the effects of segmental excluded volume in the polymer

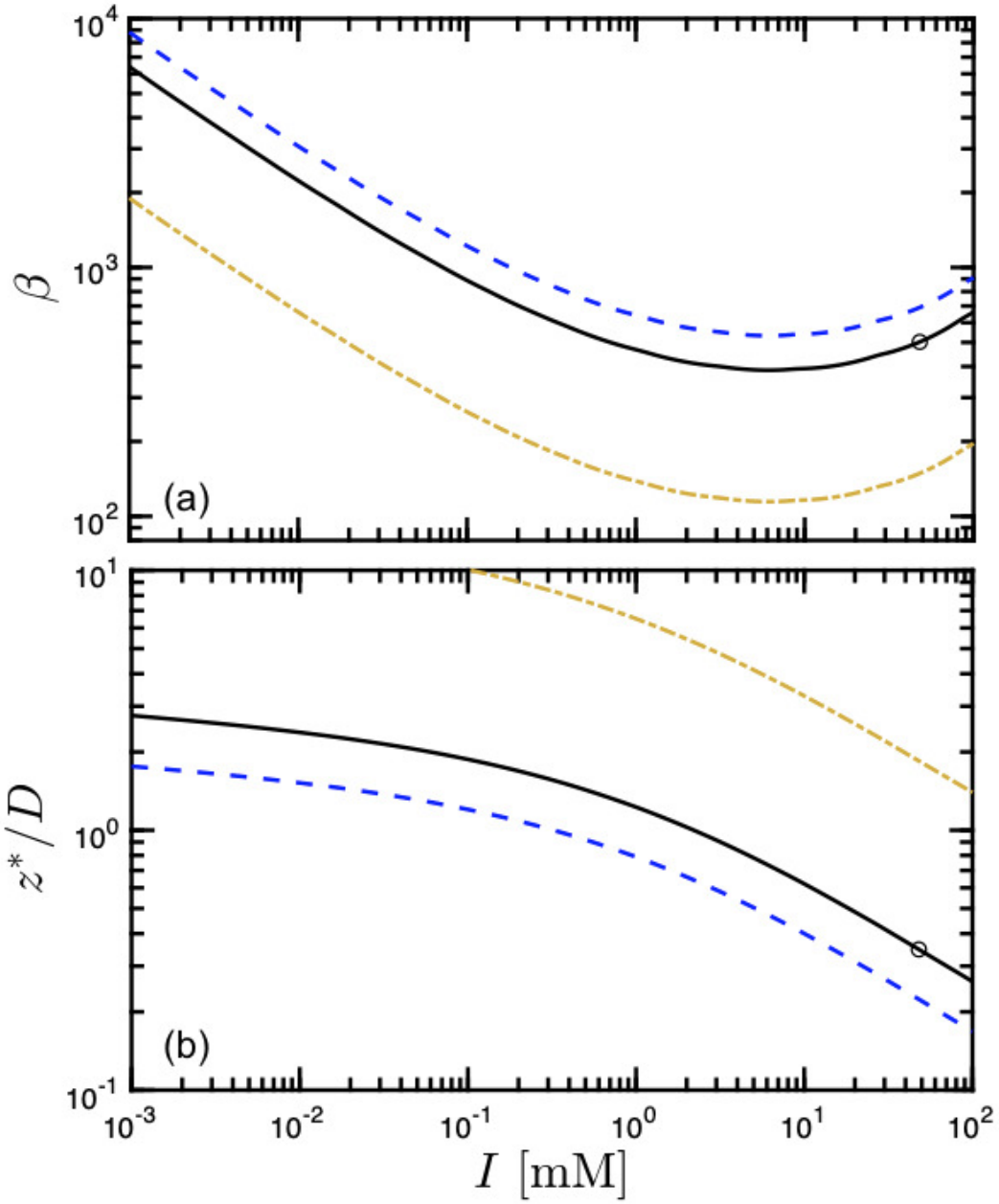

Figure 4.8: Plot of (a) the dimensionless wall interaction parameter β given by Eq. (4.15) and (b) the position z^* at which the wall interaction potential $\beta\phi(z)$ decays to $0.1k_{\mathrm{B}}T$ as a function of buffer ionic strength for channel sizes that are proximate to the Odijk regime. The dashed blue curve corresponds to $D = l_{\mathrm{p}}$, the solid black line corresponds to the ratio $D = 0.62l_{\mathrm{p}}$ for a 34 nm channel and a 48 mM ionic strength, and the dash-dot gold line corresponds to $D = 0.1l_{\mathrm{p}}$. The black circles are the results of the analysis for the experiments of §4.4. Note that the ratio z^*/D is the DNA-wall electrostatic interaction length due to a single wall relative to the entire channel size.

model and treating the electrostatic problem without any ion correlation effects. It nevertheless illuminates two challenging directions for improving our understanding of DNA confinement in such small channels. At a fundamental level, it is worthwhile to examine how the Odijk theory is affected by polymer-wall interactions that are long-ranged. There is already evidence [116] that the mapping to an effective channel size fails for sufficiently strong wall interactions. A more detailed analysis is possible by solving Eq. (4.10) for a model potential, following previous work on confinement of ideal chains [133], or by simulating the discrete wormlike chain model for small channel sizes [116]. Such an analysis would also benefit by considering in detail the channel shape, rather than the simplified one-dimensional model used here. Moving past this relatively straightforward question poses significant challenges for theory and simulation, such as providing a more realistic electrostatic model that includes electrostatic correlations that would affect the DNA configuration [139] and capturing the physical chemistry of the interactions between different ions and DNA [140, 141].

4.6 Concluding Remarks

While the confined neutral wormlike chain model has proven to be a robust model for describing DNA confinement in relatively wide nanochannels [80–82], where the electrostatic interactions are localized to a small region near the channel walls, the results presented here suggest that these interactions play an important role in governing the extension of DNA for channels close the persistence length. This conclusion confers both challenges and opportunities. Developing a detailed model for the various electrostatic interactions taking place between a highly charged polyelectrolyte confined in close proximity to charged walls is not trivial, especially if ion correlation effects and the physicochemical details of the DNA and the particular ions play an important role. However, if such a polyelectrolyte model were available, it could be used to great benefit to understand the strong extension of DNA in nanochannels smaller than its persistence length, which underlies the genome mapping technology [29] and, as our analysis suggests, remains an unsolved problem. We are optimistic that such a model could also be merged with the basic principles underlying the telegraph model since the idea of projecting the three-dimensional walk of the DNA onto an effective one-dimensional walk of

the telegraph particle should be robust to the details of the walk that is being projected. In particular, analysis of the parameters in the telegraph model reveals that the scaling parameter α is most sensitive to the global persistence length [4], which itself is a very sensitive function of the channel size (for hard walls) [51, 129]. Understanding how the global persistence length is affected by DNA-wall electrostatic interactions is the most promising route towards improving the agreement between theory and experiment for channel-confined DNA. Importantly, the global persistence length is a property of ideal wormlike chains [129], and there exists a powerful propagator approach [142] for computing it. Modifying that approach [142] to account for electrostatic interactions, even within the simple models used here, provides an enticing opportunity.

Our analysis of DNA-wall interactions also provides an interesting perspective on different methods for equilibrium stretching of DNA by channel confinement. Two general approaches have been proposed, the relatively high ionic strength, small channel approach used here [29] and a low ionic strength, large channel approach [143, 144]. To date, the metric used to compare methods is the ratio of the channel size to persistence length, D/l_{p}, since it is the relevant parameter for the Odijk theory in Eq. (4.13). Our analysis suggest that the DNA extension should be a function of both D/l_{p} and the wall interaction parameter β in Eq. (4.15). A particularly strong test of our conclusions would be to examine these two approaches at the same values of $D/l_{\mathrm{p}} < 1$. The electrostatic interactions embodied in β, as well as the length scale describing their decay, will be different for the two systems due to the nonlinearity of the electrostatics models. If these experiments furnished different DNA extensions at the same value of the Odijk parameter D/l_{p}, this would be a strong indication that DNA-wall electrostatic effects indeed play a prominent role.

Chapter 5

Conclusion and Discussion

As stated in the beginning, understanding the behavior of DNA in confinement and the corresponding physical properties is like a huge puzzle. We aimed to find the missing pieces of the puzzle to give a more comprehensive model for DNA in confinement when interpreting experimental results. Along the way of the work, some interesting questions came out and are worthwhile to be further studied. As such, we are going to review the important accomplishments in this dissertation, and discuss future research direction from these open questions with some preliminary results in this chapter.

The behavior of DNA in confinement can be roughly differentiated as four regimes as a function of channel size. These four regimes can be regarded as the edge pieces of the puzzle. The frame of the puzzle began with the corner pieces (the Odijk regime and the de Gennes regime), and was gradually connected with the edge pieces by the development of theories within the two limiting cases (the backfolded Odijk regime and the extended de Gennes regime). The key physical parameters used to characterize DNA are like the unifying aspects of the puzzle. Those are the obvious features which are much like the easy regions that make up the body of a puzzle. While one would think the puzzle is about to be completed, it turns out there are missing pieces which are crucial to connecting multiple features or between the feature and the edge. This dissertation work contains two missing pieces of the puzzle. The most important findings are that there are couples of oversimplified assumptions which were misused or even not considered in previous studies. In addition, this research work provided convincing conclusions based on high-throughput data, generated by a genome mapping technique,

and relevant statistical analysis, which highlights the importance of combining different disciplines in science to achieve research goals.

While the sequence-dependent bending energy of short DNA molecules has been studied for many decades, how this behavior is connected to the persistence length of DNA at long length scales remains unclear. For simplicity, an approximation of sequence-independent persistence length is usually used to interpret experimental results when using long DNA as a model polymer for physics experiments. In Chapter 3, we tested this conventional assumption by measuring the extension of human DNA in nanochannels with a high-throughput genome mapping technique. Around 20% increase in the persistence length was found as % GC content increases. The key outcome of the work is a new model that greatly captures the experimental results. This model, which contains a sequence-dependent intrinsic persistence length and a sequence-independent electrostatic persistence length, is straightforward to use in the analysis of experimental data.

While DNA is taken as a neutral polymer in theories for the behavior of DNA in confinement, a comparison between a series of experimental data measured at different channel sizes and theory showed a breakdown in the model as the channel size approaches the Odijk regime of strong confinement. The result revealed that a systematic source of error exists in the whole system that cannot be resolved by a simple adjustment in the calculation process. A dimensional analysis later showed the DNA-wall electrostatic interactions which were ignored in the model might be the source of disagreement. This discovery provides a potential explanation to the discrepancy in the previous studies [80,83] between experiments and simulation or experiments and theory.

Based on the findings in Chapter 3 and 4, we observe that the local properties of DNA such as the persistence length and the electrostatic forces play an important role when characterizing the overall behavior of DNA in confinement. It is reasonable to then postulate the following: is the electrostatic persistence length also dependent on the sequence? If the DNA-wall electrostatic interactions do depend on the sequence and will have significant effects at small regions, the electrostatic persistence length should also be dependent on the DNA sequence in the same manner. To begin, we examine how the ions in solution interact with DNA. At low salt concentration, where there are fewer cations in solution, DNA molecules tend to stretch more in confinement as

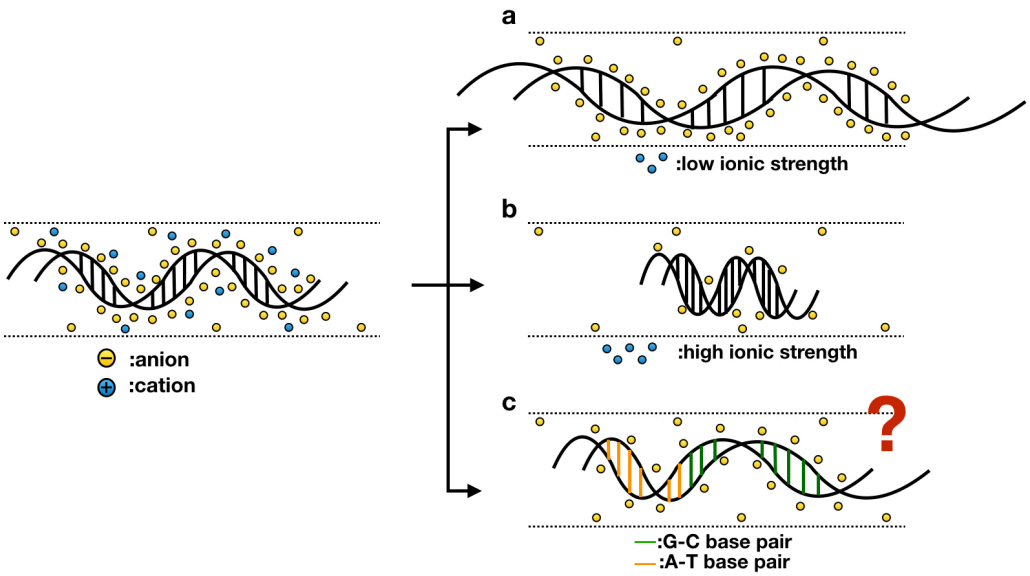

Figure 5.1: Schematic of DNA-ions electrostatics caused by different ionic strength. (a) DNA stretches more at low ionic strength due to fewer cations which reduces the screening of backbone charges. (b) DNA stretches less at high ionic strength due to more cations which increases the screening of backbone charges. (c) An idea of whether DNA-ions electrostatic interactions depend on DNA sequences due to different numbers of highly electronegative atoms in G-C and A-T base pairs.

shown in Fig. 5.1a. One of the major assumption of the model in Chapter 3 is that all sequences are affected by DNA-ions electrostatic interactions in the same manner, so the electrostatic persistence length is kept the same albeit varying % GC content at some particular ionic strength in the original experiment. This choice was based on the assumption of exclusive binding between cations and phosphate groups. However, there is literature indicating that there are more electronegative atoms in G-C base pairs than that in A-T base pairs [145], which means sequences with GC-rich region might bind a different number of cations than AT-rich region as shown in Fig. 5.1c. Uneven distribution of cation density near the minor groove [140, 141] might also affect the ion-DNA electrostatic interactions and further influence the overall behavior of DNA in confinement.

In fact, we did test the idea by measuring the extension of human DNA in a 30 nm $\times$ 30 nm nanochannel at various ionic strength with the same experimental method

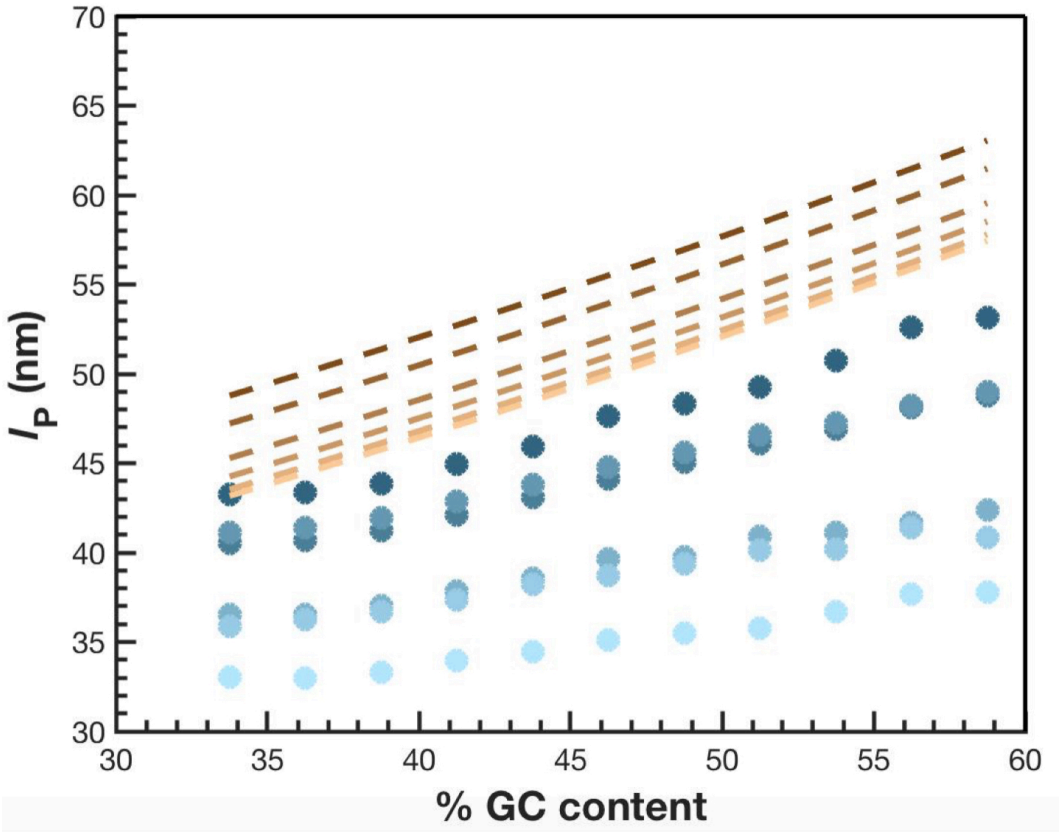

Figure 5.2: Persistence length as a function of % GC content at various ionic strength. Circles are experimental data using $D_{\text{eff}} = D - w$ at each ionic strength. Dashed lines are the statistical terpolymer model prediction in Eq. (3.13). The ionic strength of each curve of circles/dashed lines is 100 mM, 89 mM, 72 mM, 55 mM, 36 mM, and 27mM from bottom to top, respectively.

described in Ref. [83], and the preliminary result is shown in Fig. 5.2. The prediction lines (dashed lines) deviate significantly from the experimental curves (circles), with slopes increasing as the ionic strength decreases. Even if the deviation of intercept between the prediction lines and experimental results can be attributed to a systematic error of how salt concentration was changed in the experimental method, the reason for the increase in gradient as the ionic strength decreases and how the model can be modified to account for that remain unclear.

Fortunately, statistical procedures might help answer these questions. Some techniques such as principal component analysis (PCA) and regression can help in filtering

out the most relevant variables from a bunch of correlated variables to make a predictive model using an appropriate algorithm. As mentioned in chapter 3, a question like how the intrinsic curvatures affecting the overall behavior of long DNA in confinement, has multiple correlated variables. Sequences such as poly(A) tracts, $\text{poly(A)}_n{\cdot}\text{poly(T)}_n$, and GGGCCC motifs that possess substantial intrinsic curvature [112, 118–122] may have effects of different extents on the behavior of long DNA. How to quantify the effects of each candidate sequence requires a huge database and an effective calculating approach. In fact, the use of statistical methods alongside an abundance of data makes up a powerful field of research, machine learning, which is one of the most popular scientific studies in these decades. It might be useful to employ machine learning insights to the research of polymer physics. With the huge amount of genomic data at hand and the application of machine learning, the question raised above can be, optimistically, solved in a short time.

In the end, we believe that the pieces of puzzles that we found along the way and those in the near future will provide a firm foundation for the analysis of DNA-based experiments to improve genome mapping technologies.

Bibliography

[1] W. F. Reinhart, J. G. Reifenberger, D. Gupta, A. Muralidhar, J. Sheats, H. Cao, and K. D. Dorfman. Distribution of distances between DNA barcode labels in nanochannels close to the persistence length. *J. Chem. Phys.*, 142:064902, 2015.

[2] S. Geggier and A. Vologodskii. Sequence dependence of DNA bending rigidity. *Proc. Natl. Acad. Sci. USA*, 107(35):15421–15426, 2010.

[3] A. B. Bhandari and K. D. Dorfman. Simulations corroborate telegraph model predictions for the extension distributions of nanochannel confined DNA. *Biomicrofluidics*, 13(4):044110, 2019.

[4] E. Werner, G. K. Cheong, D. Gupta, K. D. Dorfman, and B. Mehlig. One-parameter theory for DNA extension in a nanochannel. *Phys. Rev. Lett.*, 119:268102, 2017.

[5] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[6] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561, 1970.

[7] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, and R. A. Holt. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[8] T. C. Sequencing, R. H. Waterson, E. S. Lander, R. K. Wilson, and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69, 2005.

[9] A. Fujiyama, H. Watanabe, A. Toyoda, T. D. Taylor, T. Itoh, S.-F. Tsai, H.-S. Park, M.-L. Yaspo, H. Lehrach, and Z. Chen. Construction and analysis of a human-chimpanzee comparative clone map. *Science*, 295(5552):131–134, 2002.

[10] DNA. `https://en.wikipedia.org/wiki/DNA`.

[11] The cost of sequencing a human genome. `https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost`.

[12] R. Wu and E. Taylor. Nucleotide sequence analysis of DNA: II. Complete nucleotide sequence of the cohesive ends of bacteriophage λ DNA. *J. Mol. Biol.*, 57(3):491–511, 1971.

[13] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*, 74(12):5463–5467, 1977.

[14] E. R. Mardis. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402, 2008.

[15] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA*, 74(2):560–564, 1977.

[16] O. Morozova and M. A. Marra. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–264, 2008.

[17] K. J. Hoff. The effect of sequencing errors on metagenomic gene prediction. *BMC Genom.*, 10(1):520, 2009.

[18] F. Sanger, A. R. Coulson, B. G. Barrell, A. J. H. Smith, and B. A. Roe. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *J. Mol. Biol.*, 143(2):161–178, 1980.

[19] B. Meder, J. Haas, A. Keller, C. Heid, S. Just, A. Borries, V. Boisguerin, M. Scharfenberger-Schmeer, P. Stähler, and M. Beier. Targeted next-generation sequencing for the molecular genetic diagnostics of cardiomyopathies. *Circ. Cardiovasc. Genet.*, 4(2):110–122, 2011.

[20] S. Behjati and P. S. Tarpey. What is next generation sequencing? *Arch. Dis. Child.*, 98(6):236–238, 2013.

[21] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. Comparison of next-generation sequencing systems. *Biomed Res. Int.*, 2012, 2012.

[22] S. C. Schuster. Next-generation sequencing transforms today's biology. *Nat. Methods*, 5(1):16, 2007.

[23] M. L. Metzker. Sequencing technologiesthe next generation. *Nat. Rev. Genet.*, 11(1):31, 2010.

[24] E. R. Mardis. A decades perspective on DNA sequencing technology. *Nature*, 470(7333):198, 2011.

[25] E. E. Eichler, R. A. Clark, and X. She. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat. Rev. Genet.*, 5(5):345, 2004.

[26] Overview of the different types of structural variations. `http://www.vce.bioninja.com.au/aos-3-heredity/molecular-genetics/mutations.html`.

[27] L. Feuk, A. R. Carson, and S. W. Scherer. Structural variation in the human genome. *Nat. Rev. Genet.*, 7(2):85, 2006.

[28] W. Cai, H. Aburatani, V. P. Stanton, D. E. Housman, Y.-K. Wang, and D. C. Schwartz. Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces. *Proc. Natl. Acad. Sci. USA*, 92(11):5164–5168, 1995.

[29] E. T. Lam, A. Hastie, C. Lin, D. Ehrlich, S. K. Das, M. D. Austin, P. Deshpande, H. Cao, N. Nagarajan, M. Xiao, and P. Y. Kwok. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.*, 30:771–776, 2012.

[30] A. Gupta, K. L. Kounovsky-Shafer, P. Ravindran, and D. C. Schwartz. Optical mapping and nanocoding approaches to whole-genome analysis. *Microfluid. Nanofluid.*, 20(3):44, 2016.

[31] S. K. Das, M. D. Austin, M. C. Akana, P. Deshpande, H. Cao, and M. Xiao. Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucl. Acid. Res.*, 38(18):e177–e177, 2010.

[32] A. Gutjahr and S.-Y. Xu. Engineering nicking enzymes that preferentially nick 5-methylcytosine-modified DNA. *Nucl. Acid. Res.*, 42(9):e77–e77, 2014.

[33] R. K. Neely, J. Deen, and J. Hofkens. Optical mapping of DNA: Single-molecule-based methods for mapping genomes. *Biopolymers*, 95(5):298–311, 2011.

[34] P. Zhang, P. H.-M. Too, J. C. Samuelson, S.-H. Chan, T. Vincze, S. Doucette, S. Bäckström, K. D. Potamousis, T. M. Schramm, and D. Forrest. Engineering bspqi nicking enzymes and application of n. bspqi in DNA labeling and production of single-strand DNA. *Protein Expr. Purif.*, 69(2):226–234, 2010.

[35] M. Levy-Sakin and Y. Ebenstein. Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy. *Curr. Opin. Biotech.*, 24(4):690–698, 2013.

[36] J. Jing, J. Reed, J. Huang, X. Hu, V. Clarke, J. Edington, D. Housman, T. S. Anantharaman, E. J. Huff, and B. Mishra. Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules. *Proc. Natl. Acad. Sci. USA*, 95(14):8046–8051, 1998.

[37] A. Bensimon, A. Simon, A. Chiffaudel, V. Croquette, F. Heslot, and D. Bensimon. Alignment and sensitive detection of DNA by a moving interface. *Science*, 265(5181):2096–2098, 1994.

[38] K. D Dorfman, S. B. King, D. W. Olson, J. D. P. Thomas, and D. R. Tree. Beyond gel electrophoresis : Microfluidic separations, fluorescence burst analysis, and DNA stretching. *Chem. Rev.*, 113:2584–2667, 2013.

[39] E. Y. Chan, N. M. Goncalves, R. A. Haeusler, A. J. Hatch, J. W. Larson, A. M. Maletta, G. R. Yantz, E. D. Carstea, M. Fuchs, and G. G. Wong. DNA mapping using microfluidic stretching and single-molecule detection of fluorescent site-specific tags. *Genome Res.*, 14(6):1137–1146, 2004.

[40] K. M. Phillips, J. W. Larson, G. R. Yantz, C. M. D'Antoni, M. V. Gallo, K. A. Gillis, N. M. Goncalves, L. A. Neely, S. R. Gullans, and R. Gilmanshin. Application of single molecule technology to rapidly map long DNA and study the conformation of stretched DNA. *Nucl. Acid. Res.*, 33(18):5829–5837, 2005.

[41] J. W. Larson, G. R. Yantz, Q. Zhong, R. Charnas, C. M. DAntoni, M. V. Gallo, K. A. Gillis, L. A. Neely, K. M. Phillips, and G. G. Wong. Single DNA molecule stretching in sudden mixed shear and elongational microflows. *Lab Chip*, 6(9):1187–1199, 2006.

[42] J. M. Kim and P. S. Doyle. Design and numerical simulation of a DNA electrophoretic stretching device. *Lab Chip*, 7(2):213–225, 2007.

[43] Schematic of DNA in confinement at different scales. `https://bionanogenomics.com`.

[44] W. Reisner, J. N. Pedersen, and R. H. Austin. DNA confinement in nanochannels: physics and biological applications. *Rep. Prog. Phys.*, 75:106601, 2012.

[45] C.-C. Hsieh, A. Balducci, and P. S. Doyle. An experimental study of DNA rotational relaxation time in nanoslits. *Macromolecules*, 40(14):5196–5205, 2007.

[46] J. M. Zook, D. Catoe, J. McDaniel, L. Vang, N. Spies, A. Sidow, Z. Weng, Y. Liu, C. E. Mason, N. Alexander, E. Henaff, A. B. R. McIntyre, D. Chandramohan, F. Chen, E. Jaeger, A. Moshrefi, K. Pham, W. Stedman, T. Liang, M. Saghbini, Z. Dzakula, A. Hastie, H. Cao, G. Deikus, E. Schadt, R. Sebra, A. Bashir, R. M. Truty, C. C. Chang, N. Gulbahce, K. Zhao, S. Ghosh, F. Hyland, Y. Fu, M. Chaisson, C. Xiao, J. Trow, S. T. Sherry, A. W. Zaranek, M. Ball, J. Bobe, P. Estep, G. M. Church, P. Marks, S. Kyriazopoulou-Panagiotopoulou, G. X. Y. Zheng, M. Schnall-Levin, H. S. Ordonez, P. A. Mudivarti, K. Giorda, Y. Sheng, K. B. Rypdal, and M. Salit. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data*, 3:160025, 2016.

[47] A. C. Y. Mak, Y. Y. Y. Lai, E. T. Lam, T. P. Kwok, A. K. Y. Leung, A. Poon, Y. Mostovoy, A. R. Hastie, W. Stedman, T. Anantharaman, W. Andrews, X. Zhou, A. W. C. Pang, H. Dai, C. Chu, C. Lin, J. J. K. Wu, C. M. L.

Li, J. W. Li, A. K. Y. Yim, S. Chan, J. Sibert, Z. Dzakula, H. Cao, S. M. Yiu, T. F. Chan, K. Y. Yip, M. Xiao, and P. Y. Kwok. Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics*, 202:351–362, 2016.

[48] M. Daoud and P. G. de Gennes. Statistics of macromolecular solutions trapped in small pores. *J. Phys.*, 38:85–93, 1977.

[49] T. Odijk. Scaling theory of DNA confined in nanochannels and nanoslits. *Phys. Rev. E*, 77:060901(R), 2008.

[50] L. Dai, J. Van Der Maarel, and P. S. Doyle. Extended de Gennes regime of DNA confined in a nanochannel. *Macromolecules*, 47:2445–2450, 2014.

[51] A. Muralidhar, D. R. Tree, and K. D. Dorfman. Backfolding of wormlike chains confined in nanochannels. *Macromolecules*, 47:8446–8458, 2014.

[52] J. Skolnick and M. Fixman. Electrostatic persistence length of a wormlike polyelectrolyte. *Macromolecules*, 10:944–948, 1977.

[53] T. Odijk. Polyelectrolytes near the rod limit. *J. Polym. Sci. Polym. Phys. Ed.*, 15:477–483, 1977.

[54] R. Borsali, H. Nguyen, and R. Pecora. Small-angle neutron scattering and dynamic light scattering from a polyelectrolyte solution: DNA. *Macromolecules*, 31(5):1548–1555, 1998.

[55] C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith. Entropic elasticity of λ-phage DNA. *Science*, 265(5178):1599–1600, 1994.

[56] C. Rivetti, M. Guthold, and C. Bustamante. Scanning force microscopy of DNA deposited onto mica: Equilibration versus kinetic trapping studied by statistical polymer chain analysis. *J. Mol. Biol.*, 264:919–932, 1996.

[57] E. S. G. Shaqfeh. The dynamics of single-molecule DNA in flow. *J. Non-Newton Fluid.*, 130(1):1–28, 2005.

[58] F. Latinwo and C. M. Schroeder. Model systems for single molecule polymer dynamics. *Soft Matter*, 7(18):7907–7913, 2011.

[59] A. D. Bates and A. Maxwell. *DNA topology*. Oxford University Press, USA, 2005.

[60] K. Günther, M. Mertig, and R. Seidel. Mechanical and structural properties of YOYO-1 complexed DNA. *Nucl. Acid. Res.*, 38(19):6526–6532, 2010.

[61] C. U. Murade, V. Subramaniam, C. Otto, and M. L. Bennink. Force spectroscopy and fluorescence microscopy of dsDNA-YOYO-1 complexes: implications for the structure of dsDNA in the overstretching region. *Nucl. Acid. Res.*, 38(10):3423–3431, 2010.

[62] C. G. Baumann, S. B. Smith, V. A. Bloomfield, and C. Bustamante. Ionic effects on the elasticity of single DNA molecules. *Proc. Natl. Acad. Sci. USA*, 94(12):6185–6190, 1997.

[63] M. Doi and S. F. Edwards. *The theory of polymer dynamics*, volume 73. Oxford University Press, 1988.

[64] D. R. Tree, A. Muralidhar, P. S. Doyle, and K. D. Dorfman. Is DNA a good model polymer? *Macromolecules*, 46:8369–8382, 2013.

[65] A. R. Khokhlov. *Statistical physics of macromolecules*. Amer Inst of Physics, 1994.

[66] G. S. Manning. The critical onset of counterion condensation: A survey of its experimental and theoretical basis. *Ber. Bunsenges. Phys. Chem.*, 100(6):909–922, 1996.

[67] A. V. Dobrynin. Electrostatic persistence length of semiflexible and flexible polyelectrolytes. *Macromolecules*, 38:9304–9314, 2005.

[68] M. Hogan, J. LeGrange, and B. Austin. Dependence of DNA helix flexibility on base composition. *Nature*, 304:752–754, 1983.

[69] B. Audit, C. Vaillant, A. Arneodo, Y. d'Aubenton-Carafa, and C. Thermes. Long-range correlations between DNA bending sites: relation to the structure and dynamics of nucleosomes. *J. Mol. Biol.*, 316(4):903–918, 2002.

[70] T. J. Richmond and C. A. Davey. The structure of DNA in the nucleosome core. *Nature*, 423(6936):145–150, 2003.

[71] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, 2006.

[72] G. S. Freeman, J. P. Lequieu, D. M. Hinckley, J. K. Whitmer, and J. J. de Pablo. DNA shape dominates sequence affinity in nucleosome formation. *Phys. Rev. Lett.*, 113:168101, 2014.

[73] L. J. Fetters, D. J. Lohse, T. D. Richter, T. A. Witten, and A. Zirkelt. Connection between polymer molecular weight, density, chain dimensions, and melt viscoelastic properties. *Macromolecules*, 27(17):4639–4647, 1994.

[74] T. J. Thomas and V. A. Bloomfield. Chain flexibility and hydrodynamics of the B and Z forms of poly(dG-dC)·poly(dG-dC). *Nucl. Acid. Res.*, 11(6):1919–1930, 1983.

[75] D. Stigter. Interactions of highly charged colloidal cylinders with applications to double-stranded DNA. *Biopolymers*, 16:1435–1448, 1977.

[76] T. Nicolai and M. Mandel. The ionic strength dependence of the second virial coefficient of low molar mass DNA fragments in aqueous solutions. *Macromolecules*, 22(1):438–444, 1989.

[77] A. A. Brian, H. L. Frisch, and L. S. Lerman. Thermodynamics and equilibrium sedimentation analysis of the close approach of DNA molecules and a molecular ordering transition. *Biopolymers*, 20(6):1305–1328, 1981.

[78] V. V. Rybenkov, N. R. Cozzarelli, and A. V. Vologodskii. Probability of DNA knotting and the effective diameter of the DNA double helix. *Proc. Natl. Acad. Sci. USA*, 90(11):5307–5311, 1993.

[79] L. Dai, C. B. Renner, and P. S. Doyle. The polymer physics of single DNA confined in nanochannels. *Adv. Colloid Interface Sci.*, 232:80–100, 2015.

[80] D. Gupta, J. Sheats, A. Muralidhar, J. J. Miller, D. E. Huang, S. Mahshid, K. D. Dorfman, and W. Reisner. Mixed confinement regimes during equilibrium confinement spectroscopy of DNA. *J. Chem. Phys.*, 140:214901, 2014.

[81] V. Iarko, E. Werner, L. K. Nyberg, V. Müller, J. Fritzsche, T. Ambjörnsson, J. P. Beech, J. O. Tegenfeldt, K. Mehlig, F. Westerlund, and B. Mehlig. Extension of nanoconfined DNA: Quantitative comparison between experiment and theory. *Phys. Rev. E*, 92:062701, 2015.

[82] D. Gupta, J. J. Miller, A. Muralidhar, S. Mahshid, W. Reisner, and K. D. Dorfman. Experimental evidence of weak excluded volume effects for nanochannel confined DNA. *ACS Macro Lett.*, 4:759–763, 2015.

[83] A. B. Bhandari, J. G. Reifenberger, H.-M. Chuang, H. Cao, and K. D. Dorfman. Measuring the wall depletion length of nanoconfined DNA. *J. Chem. Phys.*, 149:104901, 2018.

[84] W. F. Reinhart, J. G. Reifenberger, D. Gupta, A. Muralidhar, J. Sheats, H. Cao, and K. D. Dorfman. Erratum: "Distribution of distances between DNA barcode labels in nanochannels close to the persistence length" J. Chem. Phys. 142, 064902 (2015). *J. Chem. Phys.*, 147:029901, 2017.

[85] J. Sheats, J. G. Reifenberger, H. Cao, and K. D. Dorfman. Measurements of DNA barcode label separations in nanochannels from time-series data. *Biomicrofluidics*, 9:064119, 2015.

[86] J. G. Reifenberger, K. D. Dorfman, and H. Cao. Topological events in single molecules of *E. coli* DNA confined in nanochannels. *Analyst*, 140:4887–4894, 2015.

[87] B. Kundukad, J. Yan, and P. S. Doyle. Effect of YOYO-1 on the mechanical properties of DNA. *Soft Matter*, 10(48):9721–9728, 2014.

[88] S. R. Quake, H. Babcock, and S. Chu. The dynamics of partially extended single molecules of DNA. *Nature*, 388(6638):151, 1997.

[89] T. Berge, N. S. Jenkins, R. B. Hopkirk, M. J. Waring, J. M. Edwardson, and R. M. Henderson. Structural perturbations in DNA caused by bis-intercalation of dibucalinium visualised by atomic force microscopy. *Nucl. Acid. Res.*, 30(13):2980–2986, 2002.

[90] A. Sischka, K. Toensing, R. Eckel, S. D. Wilking, N. Sewald, R. Ros, and D. Anselmetti. Molecular mechanisms and kinetics between DNA and DNA binding ligands. *Biophys. J.*, 88(1):404–411, 2005.

[91] M. Maaloum, P. Muller, and S. Harlepp. DNA-intercalator interactions: structural and physical analysis using atomic force microscopy in solution. *Soft Matter*, 9(47):11233–11240, 2013.

[92] N. Shi and V. M. Ugaz. An entropic force microscope enables nano-scale conformational probing of biomolecules. *Small*, 10(13):2553–2557, 2014.

[93] T. Odijk. On the statistics and dynamics of confined or entangled stiff polymers. *Macromolecules*, 16:1340–1344, 1983.

[94] Y. Yang, T. W. Burkhardt, and G. Gompper. Free energy and extension of a semiflexible polymer in cylindrical confining geometries. *Phys. Rev. E*, 76:011804, 2007.

[95] T. W. Burkhardt, Y. Yang, and G. Gompper. Fluctuations of a long, semiflexible polymer in a narrow channel. *Phys. Rev. E*, 82:041801, 2010.

[96] W. Reisner, K. J. Morton, R. Riehn, Y. M. Wang, Z. Yu, M. Rosen, J. C. Sturm, S. Y. Chou, E. Frey, and R. H. Austin. Statics and dynamics of single DNA molecules confined in nanochannels. *Phys. Rev. Lett.*, 94:196101, 2005.

[97] C. Zhang, F. Zhang, J. A. van Kan, and J. R. C. van der Maarel. Effects of electrostatic screening on the conformation of single DNA molecules confined in a nanochannel. *J. Chem. Phys.*, 128:225109, 2008.

[98] F. Persson, P. Utko, W. Reisner, N. B. Larsen, and A. Kristensen. Confinement spectroscopy: Probing single DNA molecules with tapered nanochannels. *Nano Lett.*, 9:1382–1385, 2009.

[99] P. Cifra. Channel confinement of flexible and semiflexible macromolecules. *J. Chem. Phys.*, 131:224903, 2009.

[100] Y. Wang, D. R. Tree, and K. D. Dorfman. Simulation of DNA extension in nanochannels. *Macromolecules*, 44:6594–6604, 2011.

[101] D. R. Tree, Y. Wang, and K. D. Dorfman. Extension of DNA in a nanochannel as a rod-to-coil transition. *Phys. Rev. Lett.*, 110:208103, 2013.

[102] J. Z. Y. Chen. Self-avoiding wormlike chain confined in a cylindrical tube: scaling behavior. *Phys. Rev. Lett.*, 121:037801, 2018.

[103] H.-M. Chuang, J. G. Reifenberger, H. Cao, and K. D. Dorfman. Sequence-dependent persistence length of long DNA. *Phys. Rev. Lett.*, 119:227802, 2017.

[104] T. T. Perkins, D. E. Smith, and S. Chu. Direct observation of tube-like motion of a single polymer chain. *Science*, 264(5160):819–822, 1994.

[105] T. T. Perkins, D. E. Smith, and S. Chu. Single polymer dynamics in an elongational flow. *Science*, 276(5321):2016–2021, 1997.

[106] D. E. Smith, T. T. Perkins, and S. Chu. Self-diffusion of an entangled DNA molecule by reptation. *Phys. Rev. Lett.*, 75(22):4146–4149, 1995.

[107] J. O. Tegenfeldt, C. Prinz, H. Cao, S. Chou, W. W. Reisner, R. Riehn, Y. M. Wang, E. C. Cox, J. C. Sturm, P. Silberzan, and R. H. Austin. The dynamics of genomic-length DNA molecules in 100-nm channels. *Proc. Natl. Acad. Sci. USA*, 101:10979–10983, 2004.

[108] K. D. Dorfman. DNA electrophoresis in microfabricated devices. *Rev. Mod. Phys.*, 82(4):2903–2947, 2010.

[109] G. S. Freeman, D. M. Hinckley, J. P. Lequieu, J. K. Whitmer, and J. J. de Pablo. Coarse-grained modeling of DNA curvature. *J. Chem. Phys.*, 141(16):165103, 2014.

[110] P. Cong, L. Dai, H. Chen, J. R. C. Van Der Maarel, P. S. Doyle, and J. Yan. Revisiting the anomalous bending elasticity of sharply bent DNA. *Biophys. J.*, 109(11):2338–2351, 2015.

[111] D. MacDonald, K. Herbert, X. Zhang, T. Polgruto, and P. Lu. Solution structure of an A-tract DNA bend. *J. Mol. Biol.*, 306(5):1081–1098, 2001.

[112] J. S. Mitchell, J. Glowacki, A. E. Grandchamp, R. S. Manning, and J. H. Maddocks. Sequence-dependent persistence lengths of DNA. *J. Chem. Theory Comput.*, 13(4):1539–1555, 2017.

[113] H. Cao, A. R. Hastie, D. Cao, E. T. Lam, Y. Sun, H. Huang, X. Liu, L. Lin, W. Andrews, S. Chan, S. Huang, X. Tong, M. Requa, T. Anantharaman, A. Krogh, H. Yang, H. Cao, and X. Xu. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *Gigascience*, 3(1):34, 2014.

[114] S. Laib, R. M. Robertson, and D. E. Smith. Preparation and characterization of a set of linear DNA molecules for polymer physics and rheology studies. *Macromolecules*, 39(12):4115–4119, 2006.

[115] A. Valouev. *Shotgun optical mapping: A comprehensive statistical and computational analysis*. PhD thesis, University of Southern California, 2006.

[116] G. K. Cheong, X. Li, and K. D. Dorfman. Wall depletion length of a channel-confined polymer. *Phys. Rev. E*, 95:022501, 2017.

[117] P. Licinio and J. Guerra. Irreducible representation for nucleotide sequence physical properties and self-consistency of nearest-neighbor dimer sets. *Biophys. J.*, 92(6):2000–2006, 2007.

[118] C. Yoon, G. G. Privé, D. S. Goodsell, and R. E. Dickerson. Structure of an alternating B-DNA helix and its relationship to A-tract DNA. *Proc. Natl. Acad. Sci. USA*, 85(17):6332–6336, 1988.

[119] M. Dlakic and R. E. Harrington. Bending and torsional flexibility of G/C-rich sequences as determined by cyclization assays. *J. Biol. Chem.*, 270(50):29945–29952, 1995.

[120] I. Brukner, S. Susic, M. Dlakic, A. Savic, and S. Pongor. Physiological concentration of magnesium ions induces a strong macroscopic curvature in GGGCCC-containing DNA. *J. Mol. Biol.*, 236(1):26–32, 1994.

[121] A. A. Travers. The structural basis of DNA flexibility. *Philos. Trans. A. Math. Phys. Eng. Sci.*, 362(1820):1423–38, 2004.

[122] J. Bednar, P. Furrer, V. Katritch, A. Z. Stasiak, J. Dubochet, and A. Stasiak. Determination of DNA persistence length by cryo-electron microscopy. Separation of the static and dynamic contributions to the apparent persistence length of DNA. *J. Mol. Biol.*, 254(4):579–594, 1995.

[123] W. Reisner, J. P. Beech, N. B. Larsen, H. Flyvbjerg, A. Kristensen, and J. O. Tegenfeldt. Nanoconfinement-enhanced conformational response of single DNA molecules to changes in ionic environment. *Phys. Rev. Lett.*, 99:058302, 2007.

[124] C.-C. Hsieh, A. Balducci, and P. S. Doyle. Ionic effects on the equilibrium dynamics of DNA confined in nanoslits. *Nano Lett.*, 8:1683–1688, 2008.

[125] H.-M. Chuang, J. G. Reifenberger, A. B. Bhandari, and K. D. Dorfman. Extension distribution for DNA confined in a nanochannel near the Odijk regime. *J. Chem. Phys.*, 151:114903, 2019.

[126] F. Brochard-Wyart, T. Tanaka, N. Borghi, and P. G. De Gennes. Semiflexible polymers confined in soft tubes. *Langmuir*, 21:4144–4148, 2005.

[127] E. Werner and B. Mehlig. Confined polymers in the extended de Gennes regime. *Phys. Rev. E*, 90:062602, 2014.

[128] D. Ödman, E. Werner, K. D. Dorfman, C. R. Doering, and B. Mehlig. Distribution of label spacings for genome mapping in nanochannels. *Biomicrofluidics*, 12:034115, 2018.

[129] T. Odijk. DNA confined in nanochannels: Hairpin tightening by entropic depletion. *J. Chem. Phys.*, 125:204904, 2006.

[130] Y. Wang, W. F. Reinhart, D. R. Tree, and K. D. Dorfman. Resolution limit for DNA barcodes in the Odijk regime. *Biomicrofluidics*, 6:014101, 2012.

[131] M. A. Stephens. EDF statistics for goodness of fit and some comparisons. *J. Am. Stat. Assoc.*, 69:730–737, 1974.

[132] J. Z. Y. Chen. Theory of wormlike polymer chains in confinement. *Prog. Polym.Sci.*, 54-55:3–46, 2016.

[133] J. Z. Y. Chen. Free energy and extension of a wormlike chain in tube confinement. *Macromolecules*, 46:9837–9844, 2013.

[134] A. Muralidhar, D. R. Tree, Y. Wang, and K. D. Dorfman. Interplay between chain stiffness and excluded volume of semiflexible polymers confined in nanochannels. *J. Chem. Phys.*, 140:084905, 2014.

[135] D. Stigter. The charged colloidal cylinder with a gouy double layer. *J. Colloid Interface Sci.*, 53:296–306, 1975.

[136] S. H. Behrens and D. G. Grier. The charge of glass and silica surfaces. *J. Chem. Phys.*, 115:6716–6721, 2001.

[137] T. Hiemstra, J. C. M. de Witt, and W. H. van Riemsdijk. Multisite proton adsorption modeling at the solid/solution interface of (hydr)oxides: A new approach. II. Applications to various important (hydr)oxides. *J. Colloid Interface Sci.*, 133:105–117, 1989.

[138] R. K. Iler, editor. *The chemistry of silica.* John Wiley and Sons, New York, 1979.

[139] M. Muthukumar. 50th Anniversary perspective: A perspective on polyelectrolyte solutions. *Macromolecules*, 50:9528–9560, 2017.

[140] L. McFail-Isom, C. C. Sines, and L. D. Williams. DNA structure: cations in charge? *Curr. Opin. Struct. Biol.*, 9:298–304, 1999.

[141] B. Jayaram, K. A. Sharp, and B. Honig. The electrostatic potential of B-DNA. *Biopolymers*, 28:975–993, 1989.

[142] J. Z. Y. Chen. Conformational properties of a back-folding wormlike chain confined in a cylindrical tube. *Phys. Rev. Lett.*, 118:247802, 2017.

[143] K. Jo, D. M. Dhingra, T. Odijk, J. J. de Pablo, M. D. Graham, R. Runnheim, D. Forrest, and D. C. Schwartz. A single-molecule barcoding system using nanoslits for DNA analysis. *Proc. Natl. Acad. Sci. USA*, 104:2673–2678, 2007.

[144] Y. Kim, K. S. Kim, K. L. Kounovsky, R. Chang, G. Y. Jung, J. J. de Pablo, K. Jo, and D. C. Schwartz. Nanochannel confinement: DNA stretch approaching full contour length. *Lab Chip*, 11:1721–1729, 2011.

[145] I. Sissoëff, J. Grisvard, and E. Guillé. Studies on metal ions-DNA interactions: Specific behaviour of reiterative DNA sequences. *Prog. Biophys. Mol. Biol.*, 31(C):165–199, 1978.

Appendix A

Supporting Information to Chapter 3

A.1 Details of the Statistical Terpolymer Model

As noted in §3.5, the bending energy is given by the sum

$$E = \sum_{i,j} p_{ij} E_{ij}, \tag{A.1}$$

where p_{ij} is the probability of having dinucleotide pair (i, j) and $E_{i,j}$ is the bending energy for that pair.

If we denote by γ the probability of picking a S in the 5'-3' direction (strong base pairing between G and C) then $1-\gamma$ is the probability of picking a W in the 5'-3' direction (weak base pairing between A and T). Assuming random statistics, the probability of picking SS is

$$p_{\mathrm{SS}} = \gamma^2, \tag{A.2}$$

the probability of picking WW is

$$p_{\mathrm{WW}} = (1-\gamma)^2, \tag{A.3}$$

and the probability of picking either SW or WS is

$$p_{\mathrm{WS}} = p_{\mathrm{SW}} = \gamma(1-\gamma). \tag{A.4}$$

Note that these probabilities are normalized properly:

$$p_{\mathrm{SS}} + p_{\mathrm{SW}} + p_{\mathrm{WS}} + p_{\mathrm{WW}} = \gamma^2 + 2\gamma(1-\gamma) + (1-\gamma)^2 = 1 \tag{A.5}$$

The bending energies are obtained from the experimental results of Hogan *et al.* [68]. We found that the absolute values of the bending energies do not lead to a useful result, but using the relative values work well. From the table in Ref. [68], we have

$$\frac{E_{\mathrm{SW}}}{E_{\mathrm{SS}}} = \frac{1.4 \times 10^9 \text{ dyne/cm}^2}{2.9 \times 10^9 \text{ dyne/cm}^2} \tag{A.6}$$

and

$$\frac{E_{\mathrm{WW}}}{E_{\mathrm{SS}}} = \frac{0.82 \times 10^9 \text{ dyne/cm}^2}{2.9 \times 10^9 \text{ dyne/cm}^2} \tag{A.7}$$

Note that the value of 2.9×10^9 dyne/cm^2 for G-C base pairs is from the longest molecules in the data set of Hogan *et al.* [68], whereas the other data come from shorter molecules.

Using only the ratios of these bending energies leads to one free parameter in the model at this stage. We chose the free parameter to be E_{SS}, and we will remove it at a later step in the analysis. Putting everything together so far,

$$E = (1-\gamma)^2 E_{\mathrm{WW}} + 2\gamma(1-\gamma)E_{\mathrm{SW}} + \gamma^2 E_{\mathrm{SS}} \tag{A.8}$$

Dividing through by E_{SS} and substituting the numerical values gives

$$\frac{E}{E_{\mathrm{SS}}} = (1-\gamma)^2 \left(\frac{0.82}{2.9}\right) + 2\gamma(1-\gamma)\left(\frac{1.4}{2.9}\right) + \gamma^2 \tag{A.9}$$

The intrinsic persistence length (at infinite ionic strength) is [68]

$$l_{\mathrm{p},0} = \frac{EI_{\mathrm{s}}}{k_{\mathrm{B}}T} \tag{A.10}$$

where I_{s} is the surface moment of inertia and $k_{\mathrm{B}}T$ is the Boltzmann factor. We can rewrite this as

$$l_{\mathrm{p},0} = \left(\frac{E_{\mathrm{SS}} I_{\mathrm{s}}}{k_{\mathrm{B}} T}\right) \left[(1-\gamma)^2 \left(\frac{0.82}{2.9}\right) + 2\gamma(1-\gamma)\left(\frac{1.4}{2.9}\right) + \gamma^2\right] \tag{A.11}$$

From Dobrynin's theory [124], the intrinsic persistence length for λ-DNA is 46.1 nm. The GC content of this sequence is $\gamma = 0.4986$. So we have

$$46.1 \text{ nm} = \left(\frac{E_{\mathrm{SS}} I_{\mathrm{s}}}{k_{\mathrm{B}} T}\right) \left[(1-0.4986)^2 \left(\frac{0.82}{2.9}\right) + 2(0.4986)(1-0.4986)\left(\frac{1.4}{2.9}\right) + (0.4986)^2\right] \tag{A.12}$$

which reduces to

$$46.1 \text{ nm} = 0.5611 \left(\frac{E_{\mathrm{SS}} I_{\mathrm{s}}}{k_{\mathrm{B}} T}\right) \tag{A.13}$$

whereupon we remove the remaining free parameter by requiring that

$$\frac{E_{\mathrm{SS}} I_{\mathrm{s}}}{k_{\mathrm{B}} T} = 82.2 \text{ nm} \tag{A.14}$$

Using this result in Eq. (A.11) gives

$$l_{\mathrm{p},0} \text{ [nm]} = (82.2) \left[(1-\gamma)^2 \left(\frac{0.82}{2.9}\right) + 2\gamma(1-\gamma)\left(\frac{1.4}{2.9}\right) + \gamma^2\right] \tag{A.15}$$

which gives us back the result for §3.5 of Eq. (3.13),

$$l_{\mathrm{p},0} \text{ [nm]} = 23 + 33\gamma + 26\gamma^2 \tag{A.16}$$

The final result is rounded to two significant digits based on the experimental data reported in Ref. [68].

The reason for using ratios of bending energies in the data of Hogan *et al.* [68], rather than the absolute values, emerges from Eq. (A.14). If we assume a rise of 0.34 nm/bp, this results suggests that the GC persistence length is 241 bp, which is approximately half that reported by Hogan *et al.* [68].

Appendix B

Supporting Information to Chapter 4

B.1 Comparison to Experimental Data for E.Coli

The following plots (Fig. B.1-B.3) are equivalent to Fig. 4.3 in §4.3 for the other channel sizes.

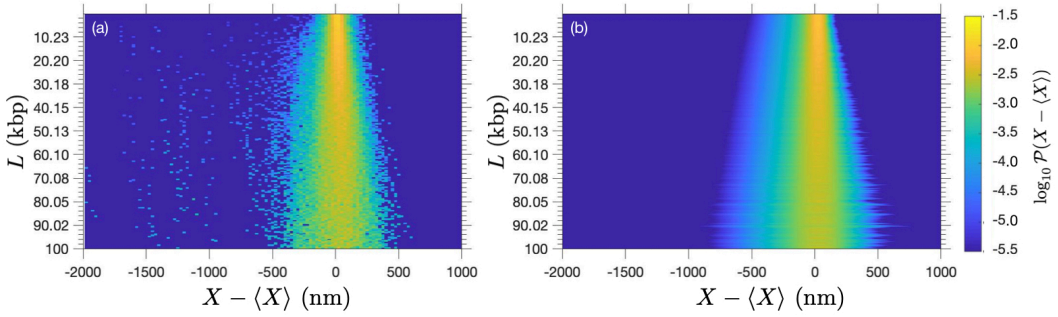

Figure B.1: Comparison between the predictions of the telegraph model in Eq. (4.6) and the experimental data of Reinhart *et al.* [1] for (a) $D = 42$ nm, experiment; and (b) $D = 42$ nm, theory.

Figure B.4 shows the comparison between the best fit value of the adjustable parameter, σ_0, and the corresponding σ_{Odijk} for each channel size D and each bin in L.

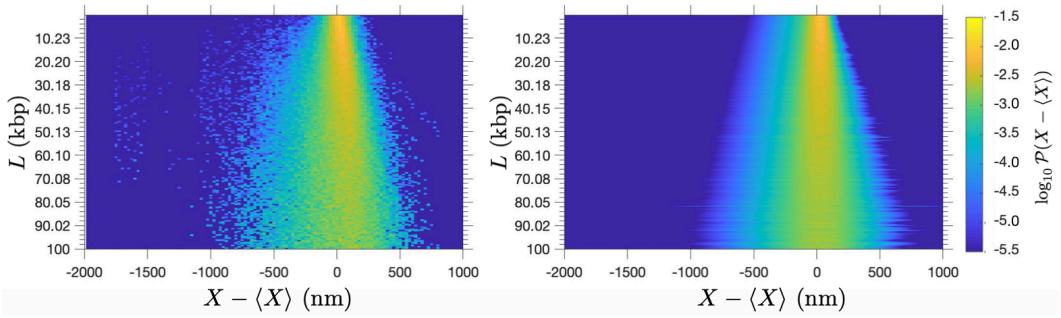

Figure B.2: Comparison between the predictions of the telegraph model in Eq. (4.6) and the experimental data of Reinhart *et al.* [1] for (a) $D = 43$ nm, experiment; and (b) $D = 43$ nm, theory.

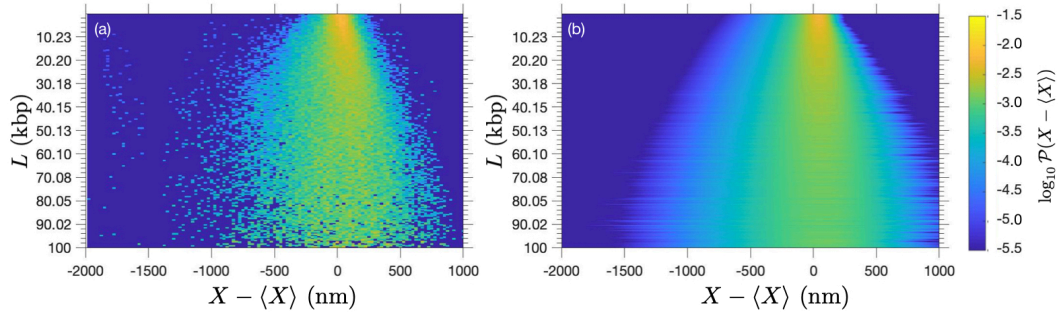

Figure B.3: Comparison between the predictions of the telegraph model in Eq. (4.6) and the experimental data of Reinhart *et al.* [1] for (a) $D = 49$ nm, experiment; and (b) $D = 49$ nm, theory.

B.2 Comparison to Experimental Data for λ-DNA

B.2.1 Additional Information for the Experimental Method of λ-DNA Experiment

All of the Possible Resolvable Labeling Patterns of λ-DNA

Figure B.5 shows all of the possible labeling patterns of λ-DNA for the correlation coefficient filter.

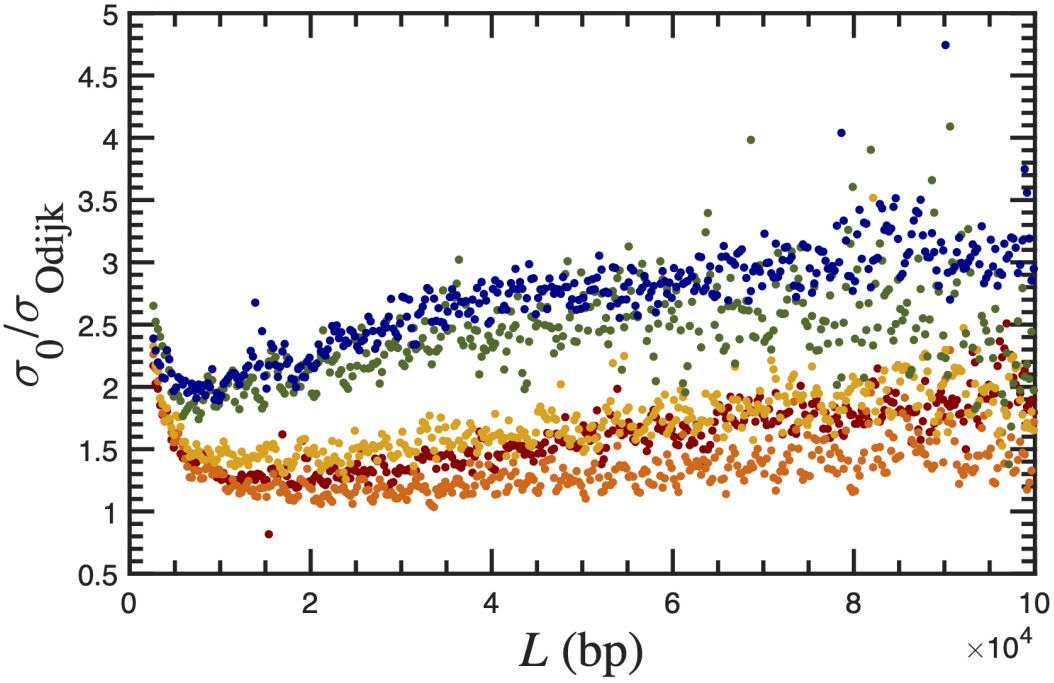

Figure B.4: The best fit value of the adjustable parameter, σ_0, obtained by fitting the right tail of distribution to the experimental data of Reinhart *et al.* [1] following the approach of Ödman *et al.*, [128] divided by the corresponding σ_{Odijk} calculated using Eq. (4.6) for each channel size. Channel sizes for the data points are 40 nm (red), 42 nm (orange), 43 nm (yellow), 49 nm (green), and 51 nm (blue), respectively.

Sensitivity Analysis to the Parameter Selection

In §4.4, to extract the intact λ-DNA molecules that were not sheared between the first and last nick sites (45,315 bp apart), we adopted a set of three filters that attenuate the effects of systematic errors in the data set. In the first filtering step, we confirmed that the molecule has the correct number of labels. A completely labeled molecule will have between 7 and 9 resolvable labels, as shown in Fig. B.5. The total number of molecuels was reduced from 1,241,907 to 612,121 in this filtering step. Next, we required that molecules have a backbone extension L between 14.0 μm and 16.5 μm. The latter value corresponds to fully extended intact λ-DNA (48,510 bp) as a conservative limiting value. The former value corresponds to the distance between the first and last nick points if the molecule were stretched to the 91.2% extent predicted by Odijk's theory [93, 95]

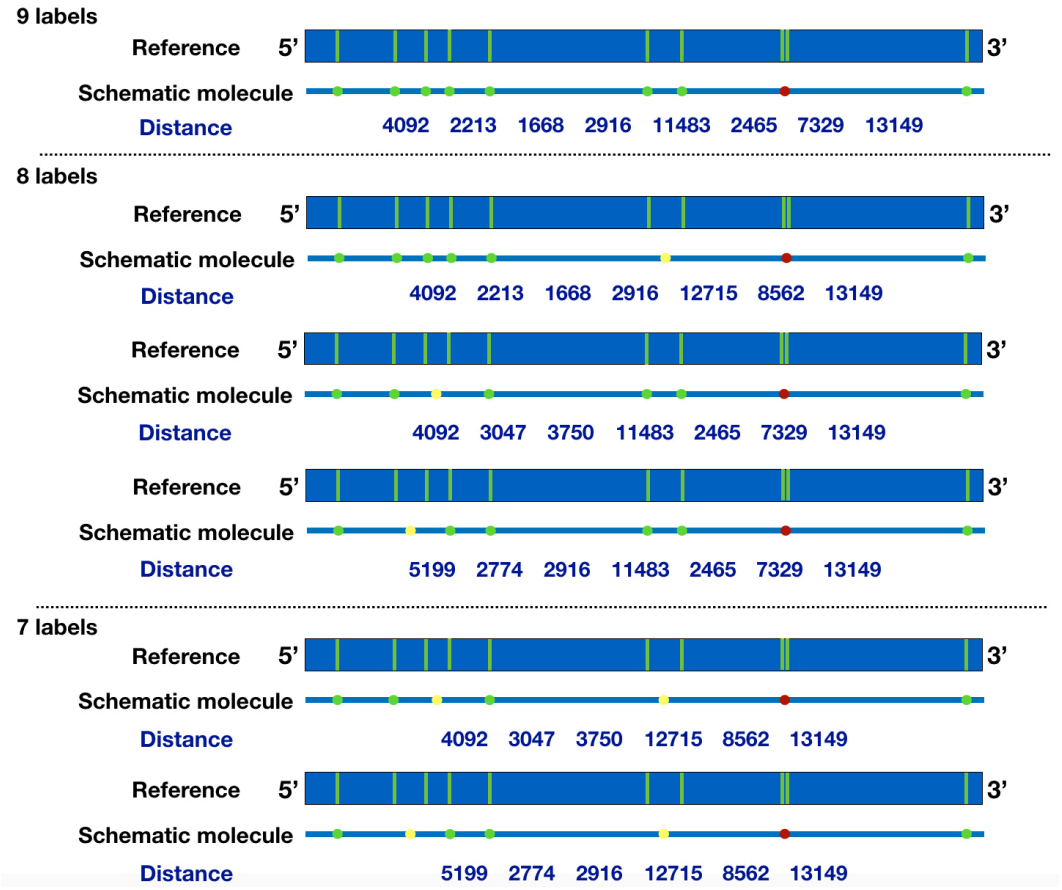

Figure B.5: All of the possible labeling patterns of λ-DNA for the correlation coefficient filter. The number refers to the distance in units of base pair between every nearest pairs of nick sites. The schematic molecules indicate the diffraction-limited spots on the reference.

assuming $D_{\text{eff}} = 26.4$ nm for a physical channel size $D = 34$ nm and an effective width $w = 7.6$ nm at an ionic strength of 48 mM, using the approximation in Eq. (4.7). After applying these lower and upper bounds to the molecule length, the number of molecules left after applying the filtering step was reduced from 612,121 to 509,070.

The parameter setting of lower and upper bounds will determine the number of molecules used for the analysis. To ensure the selection of lower and upper bounds of molecule length do not affect the conclusions of our analysis, here we applied a sensitivity analysis to the parameter selection by varying the lower and upper bounds of molecule

D_{eff} (nm)	Odijk extension (%)	lower bound (μm)	upper bound (μm)
26.4	91.2	14.0	15.0
34.0	88.6	13.7	14.6
36.5	87.8	13.5	14.5

D_{eff} (nm)	number of molecules	σ_0 (nm)
26.4	497,046	198
34.0	373,432	212
36.5	293,635	215

Table B.1: The extension, lower and upper bounds of molecule length, number of molecules left after the length filter was applied, and the best-fit parameter, σ_0, to capture the right tail of the distribution for the channel sizes considered in the sensitivity analysis. The filter for the number of labels that requires 7-9 labels for a molecule was applied in advance to the whole data set (1,241,907 molecules), and the number of molecules left after that first filter was applied is 612,121.

length. With the same setting of the filtering step in §4.4, i.e., the lower bound, 14.0 μm, and D_{eff}, 26.4 nm, we first re-considered the case where the upper bound of the intact molecule length was set at the 91.2% extent predicted by Odijk's theory [93, 95] instead of the full extension. The number of molecules is reduced by applying this adjustment to the molecule length filtering step. The corresponding distribution is shown in the green dashed line in Fig. B.6. There is nearly no difference between the two distributions when we only change the upper bound of molecule length. We next checked two other cases for D_{eff} which were discussed in §4.4, D_{eff} = 34 nm, where there are no DNA-wall excluded volume interactions for the most likely channel size, and D_{eff} = 36.5 nm, the upper bound in the estimate for the channel size [83]. Odijk's theory [93, 95] was applied to estimate the extension rate for both of the two D_{eff} cases, and the lower and the upper bounds of molecule length were adjusted accordingly. Table B.1 provides the Odijk extension, lower and upper bounds of molecule length in the unit of μm, number of molecules left after the filter was applied, and the best-fit parameter, σ_0, to capture the right tail of the distribution for each D_{eff}. The results of data distribution with D_{eff} = 34 nm, and D_{eff} = 36.5 nm are shown in yellow and red dashed lines, respectively in Fig. B.6.

From the molecules left after applying the molecular length filtering step in Table B.1, we found a large number of data points were excluded from the data set as

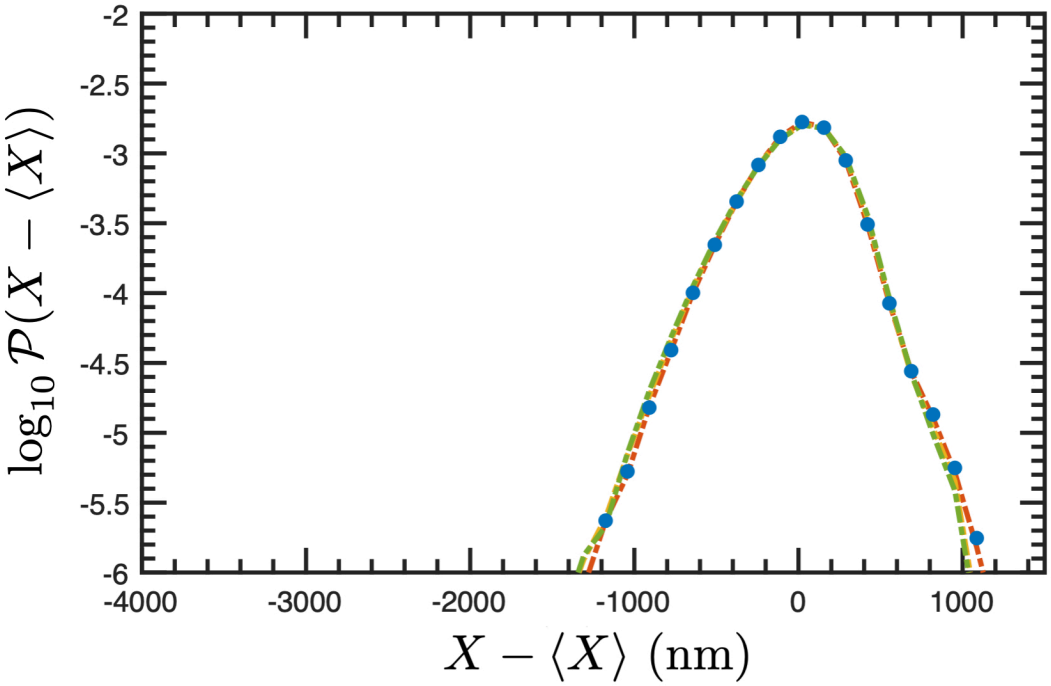

Figure B.6: Result of the data distribution with different settings for the lower and upper bounds for the molecule length following Table B.1. The effective channel sizes considered are the estimated physical channel size (34 nm, yellow dashed line), the upper bound in the physical channel size (36.5 nm, red dashed line), and the approximation in Eq. (4.7) (26.4 nm, green dashed line). The parameters of lower and upper bounds of the filter for the settings in §4.4 (blue circles) were provided in §4.4.1.

D_{eff} increased. In principle, this should increase the agreement between theory and experiment, since we would expect that these molecules represent the outliers that are broadening the experimental distribution relative to the predictions of the theory. However, the extension distribution in Fig. B.6 shows no obvious change even between the two most different cases, the one with $D_{\text{eff}} = 36.5$ nm (red dashed line) and the one used in §4.4 (blue circle). Around 40% of the data were removed by adjusting the setting of the lower and upper bounds of molecule length for these two cases, but the shapes of data distribution and the best-fit parameters, σ_0, to capture the right tail of the distribution are nearly unchanged.

We next performed a sensitivity analysis to the cutoff of correlation coefficient in

cutoff	number of molecules
0.950	298,620
0.955	273,137
0.960	241,301
0.965	216,638
0.970	202,216
0.975	189,205
0.980	166,340

Table B.2: Number of molecules left after the correlation coefficient filter was applied. Two filters were applied in advance to the whole data set (1,241,907 molecules), a filter of number of labels within 7-9 labels, and a filter of molecule length between 14.0-16.5 μm. The number of molecules left after the first two filters were applied are 612,121 and 509,070, respectively.

the last filtering step. In §4.4, we assigned a cutoff value of 0.98 to the correlation coefficient. Nearly 67% molecules from the previous filtering step were removed by this strict filter. We are interested to see if our conclusion is affected by the selection of the cutoff, which may be too stringent for filtering the data. We performed a sensitivity analysis to the parameter selection by varying the value of the correlation coefficient cutoff from 0.95 to 0.98 with 0.005 increments. The corresponding numbers of molecules left after applying this filter for each cutoff value are listed in Table B.2. Figure B.7 shows the results of data distribution for each value of cutoff.

After applying the filtering steps in Table B.2, more molecules were kept in the data set as the value of the cutoff decreased, as we would expect since we now permit less correlation between the measured label pattern and the reference. However, the distribution of extensions for the experiments in Fig. B.7 became broader as the cutoff values were lowered, which means the disagreement between the theory and the experiment is even more prominent when we relax our restriction on the cutoff of correlation coefficient.

From the sensitivity analysis to the lower and upper bounds of molecule length, we found the data distribution is unchanged and thus our conclusion was unaffected by setting a different D_{eff}. From the sensitivity analysis to the cutoff of correlation coefficient, we found our conclusion was even more convincing since including more data broadens the extension distribution when decreasing the cutoff values. We thus

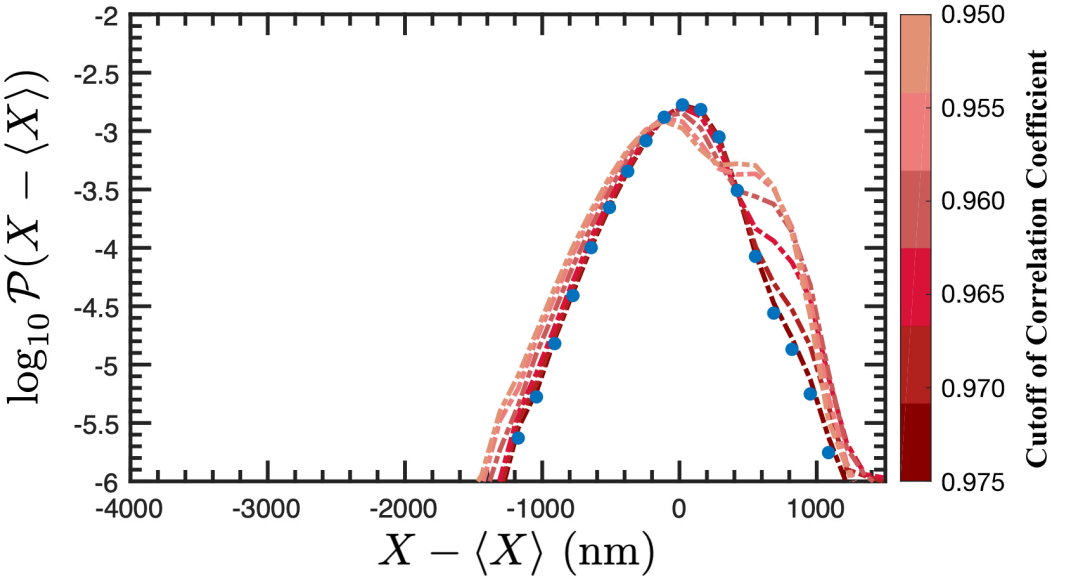

Figure B.7: Result of the data distribution for each cutoff value of correlation coefficient in Table B.2. The cutoff value of correlation coefficient used in §4.4 (blue circles) is 0.98.

conclude that our analysis is robust to the particular parameters choices used to filter the data.

B.2.2 Additional Information for the Result of λ-DNA Experiment

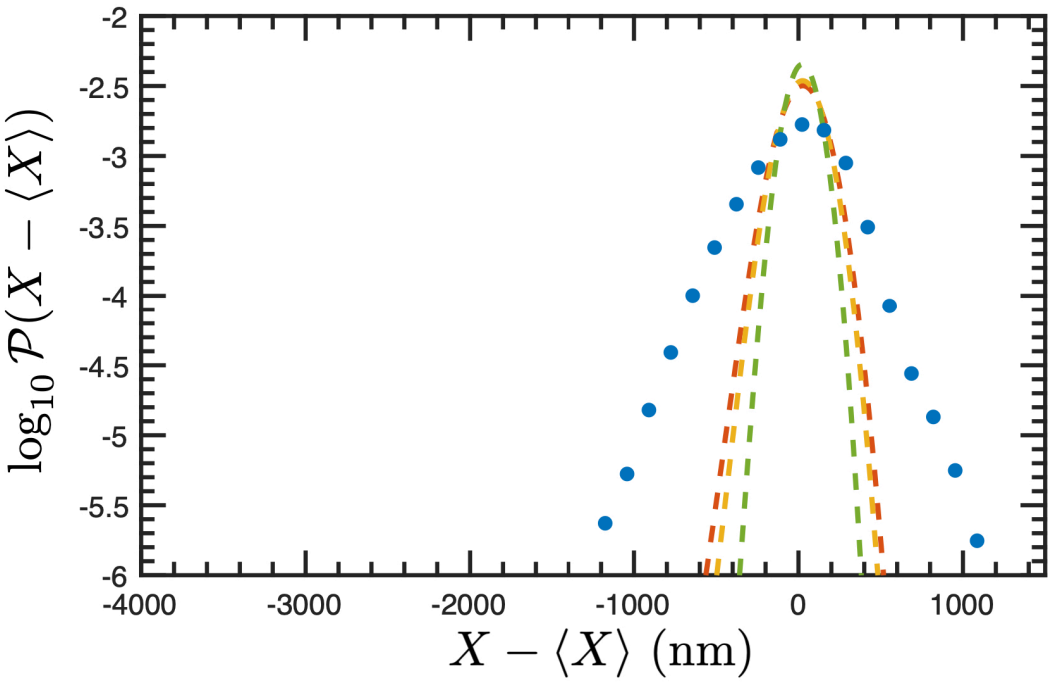

Figure B.8: Comparison between the predictions of telegraph model with $\sigma_0 = 2\sigma_{\text{Odijk}}$ (dashed lines) and the experimental data of λ-DNA (blue circle). The theoretical distributions were model at $D_{\text{eff}} = 36.5$ nm (red), $D_{\text{eff}} = 34$ nm (yellow), and $D_{\text{eff}} = 26.4$ nm (green) nanochannels using the parameter values in Table 4.2 of §4.4.2.